Rethinking Multi-domain Generalization with A General Learning Objective

Supplementary Material

7. More mathematical details of our method

7.1. Derivation details of PUB

Details for Eq. 5.

We denote $P_{mix} \triangleq \sum_n w_n P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))$. Therefore for GJSD, we have:

$$\begin{aligned}
&GJSD(\{P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))\}_{n=1}^N) \\
=&\sum\nolimits_n w_n KL(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)) \| P_{mix}) \\
=&\sum\nolimits_n w_n [H_c(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)), P_{mix}) \\
&- H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)))] \\
=&\sum\nolimits_n w_n H_c(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)), P_{mix}) \\
&- \sum\nolimits_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&\sum\nolimits_n w_n \int_{\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)} -P(x,y) \ln P_{mix}(x,y) d(x,y) \\
&- \sum\nolimits_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&\int_{\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)} -\sum\nolimits_n w_n P(x,y) \ln P_{mix}(x,y) d(x,y) \\
&- \sum\nolimits_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&\int_{\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)} -P_{mix}(x,y) \ln P_{mix}(x,y) d(x,y) \\
&- \sum\nolimits_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&H(P_{mix}) - \sum\nolimits_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))).
\end{aligned} \tag{20}$$

Therefore, the minimization of GJSD can be written as follows:

$$\begin{aligned}
&\min_{\phi,\psi} GJSD(\{P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))\}_{n=1}^N) \\
\equiv&\min_{\phi,\psi} H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) - \mathbb{E}[H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)))] \\
\equiv&\min_{\phi,\psi} H(P(\phi(\mathbf{X}), \psi(\mathbf{Y}) \mid \mathcal{D})) - \mathbb{E}[H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)))].
\end{aligned} \tag{21}$$

Details for Eq. 6. Taking into account $\mathcal{O}$, similar to [9], we have the upper bound for GJSD as:

$$\begin{aligned}
&GJSD(\{P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))\}_{n=1}^N) \\
=&H_c(P_{mix}, \mathcal{O}) - H_c(P_{mix}, \mathcal{O}) + H(P_{mix}) \\
&- \sum_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&H_c(P_{mix}, \mathcal{O}) - D_{\mathrm{KL}}(P_{mix} \| \mathcal{O}) \\
&- \sum_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
\leq&H_c(P_{mix}, \mathcal{O}) - \sum_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))).
\end{aligned} \tag{22}$$

For the standard situation where $w_1 = w_2 = ... = w_n = 1/N$, we further have:

$$\begin{aligned}
&GJSD(\{P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))\}_{n=1}^N) \\
\leq&H_c(P_{mix}, \mathcal{O}) - \sum_n w_n H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))) \\
=&H_c(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))], \mathcal{O}) \\
&- \mathbb{E}[H(P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n)))].
\end{aligned} \tag{23}$$

Details for Eq. 7. The above bound can be further reformed as:

$$\begin{aligned}
&H_c(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))], \mathcal{O}) - a \\
=&H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) \\
&+ D_{\mathrm{KL}}(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))] \| \mathcal{O}) - a, \\
=&H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) \\
&+ D_{\mathrm{KL}}(P(\phi(\mathbf{X}), \psi(\mathbf{Y})) \| \mathcal{O}) - a.
\end{aligned} \tag{24}$$

Derivation details of Eq. 8.

$$\begin{aligned}
\textbf{GAim1} =&H(P(\phi(\mathbf{X}), \psi(\mathbf{Y}) \mid \mathcal{D})) \\
=&H(P(\phi(\mathbf{X}), \psi(\mathbf{Y}), \mathcal{D})) - H(P(\mathcal{D})) \\
=&H(P(\phi(\mathbf{X}), \mathcal{D})) - H(P(\mathcal{D})) \\
&+ H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X}), \mathcal{D})) \\
=&H(P(\phi(\mathbf{X}) \mid \mathcal{D})) + H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X}), \mathcal{D})) \\
=&H(P(\psi(\mathbf{Y}) \mid \mathcal{D})) + H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y}), \mathcal{D})).
\end{aligned} \tag{25}$$

Derivation details of Eq. 9. Due to Eq. 8, we want to maintain $\phi(\mathbf{X}) \rightarrow \psi(\mathbf{Y})$ but suppressing $\psi(\mathbf{Y}) \rightarrow \phi(\mathbf{X})$. Thus we want to $\max_{\phi,\psi} H(P(\phi(\mathbf{X})|\psi(\mathbf{Y})), \mathcal{D})$ while $\min_{\phi,\psi} H(P(\psi(\mathbf{Y})|\phi(\mathbf{X})), \mathcal{D})$. which problem can

be simplified as:

$$\begin{aligned}
&\min_{\phi,\psi} H(P(\psi(\mathbf{Y})|\phi(\mathbf{X}),\mathcal{D})) - H(P(\phi(\mathbf{X})|\psi(\mathbf{Y}),\mathcal{D})) \\
=&\min_{\phi,\psi} H(\phi(\mathbf{X}),\psi(\mathbf{Y})|\mathcal{D})) - I(\phi(\mathbf{X}),\psi(\mathbf{Y})|\mathcal{D})) - \\
&H(P(\phi(\mathbf{X})|\psi(\mathbf{Y}),\mathcal{D})) + H(P(\phi(\mathbf{X})|\psi(\mathbf{Y}),\mathcal{D})) \\
=&\min_{\phi,\psi} H(\phi(\mathbf{X}),\psi(\mathbf{Y})|\mathcal{D})) - I(\phi(\mathbf{X}),\psi(\mathbf{Y})|\mathcal{D})),
\end{aligned} \tag{26}$$

where the first term is already in **GAim2**, thus **GReg2** should deal with the second term, which is:

$$\begin{aligned}
&\min_{\phi,\psi} I(\phi(\mathbf{X}),\psi(\mathbf{Y})|\mathcal{D})) = \\
&\min_{\phi,\psi} H(P(\phi(\mathbf{X}))|\mathcal{D}) - H(P(\phi(\mathbf{X}))|P(\psi(\mathbf{Y})),\mathcal{D}).
\end{aligned} \tag{27}$$

Also, due to the effect of $\mathcal{D}$ being alleviated through the mappings, the above equation is approximated as

$$\min_{\phi,\psi} H(P(\phi(\mathbf{X}))) - H(P(\phi(\mathbf{X}))|P(\psi(\mathbf{Y}))), \tag{28}$$

which is **GReg2**.

Derivation details of Eq. 10. For a Gaussian distribution $\mathcal{N}(x;\mu,\Sigma)$ with D dimension, its entropy is:

$$\begin{aligned}
H(x) =& -\int p(x)\ln p(x)dx \\
=& -\int p(x)[\ln\left((2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\right) \\
&-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)]dx \\
=&\ln\left((2\pi)^{\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\right) \\
&+\frac{1}{2}\int p(x)[(x-\mu)^\top\Sigma^{-1}(x-\mu)]dx \\
=&\ln\left((2\pi)^{\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\right)+\frac{1}{2}\int p(y)\times y^\top dy \\
=&\ln\left((2\pi)^{\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\right)+\frac{1}{2}\sum_{d=1}^{D}\mathbb{E}[y_d^2] \\
=&\ln\left((2\pi)^{\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}\right)+\frac{D}{2} \\
=&\frac{D}{2}(1+\ln(2\pi))+\frac{1}{2}\ln|\Sigma|.
\end{aligned} \tag{29}$$

Then Eq. 10 equals:

$$\begin{aligned}
&H(\mathcal{N}(\phi(\mathbf{X});\mu_{\mathbf{X}},\Sigma_{\mathbf{XX}})) \\
&-H(\mathcal{N}(\phi(\mathbf{X})\mid\psi(\mathbf{Y});\mu_{\mathbf{X}|\mathbf{Y}},\Sigma_{\mathbf{XX}|\mathbf{Y}})) \\
=&\frac{D}{2}(1+\ln(2\pi))+\frac{1}{2}\ln|\Sigma_{\mathbf{XX}}| \\
&-\frac{D}{2}(1+\ln(2\pi))-\frac{1}{2}\ln|\Sigma_{\mathbf{XX}|\mathbf{Y}}| \\
=&\frac{1}{2}\ln(|\Sigma_{\mathbf{XX}}|)-\ln(|\Sigma_{\mathbf{XX}|\mathbf{Y}}|), \\
=&\frac{1}{2}\ln(\frac{|\Sigma_{\mathbf{XX}}|}{|\Sigma_{\mathbf{XX}|\mathbf{Y}}|}).
\end{aligned} \tag{30}$$

Empirical risk. The empirical risk introduced by the whole model θ *w.r.t* $\mathbf{X},\mathbf{Y}$ is determined by a convex loss function $L(\theta)$. Following [39], the empirical risk considering $\mathcal{O}$ is:

$$\begin{aligned}
R(\theta) =& \int L(\theta)dP(\theta) + H(\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))]) \\
&+ D_{\mathrm{KL}}(\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))]\|\,\mathcal{O}) \\
&- H(P(\phi(\mathbf{X})\mid\psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X}))).
\end{aligned} \tag{31}$$

Proof of Using ψ *v.s.* not using ψ. Using *Jensen's inequality*, due to $\mathbf{Y},\psi(\mathbf{Y})$ contains the same amount of useful information as $\mathbf{Y}$, we have:

$$H(\mathbf{Y}) \geq H(\psi(\mathbf{Y})). \tag{32}$$

Therefore, we have

$$\begin{aligned}
&H(\mathbb{E}[P(\phi(\mathbf{X}_n),\mathbf{Y}_n)]) \\
=&H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) + H(\mathbb{E}[P(\mathbf{Y}_n\mid\phi(\mathbf{X}_n))]) \\
\geq&H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) + H(\mathbb{E}[P(\psi(\mathbf{Y}_n)\mid\phi(\mathbf{X}_n))]) \\
=&H(\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))]).
\end{aligned} \tag{33}$$

Therefore, for the risk of $\theta^{n\psi}$:

$$\begin{aligned}
\sup R(\theta^{n\psi}) = \sup\min_{\phi}[&\int L(\theta)dP(\theta) \\
&+ H(\mathbb{E}[P(\phi(\mathbf{X}_n),\mathbf{Y}_n)])] + b,
\end{aligned} \tag{34}$$

and the risk of $\theta^{n\psi}$:

$$\begin{aligned}
\sup R(\theta^{\psi}) = \sup\min_{\phi}[&\int L(\theta)dP(\theta) \\
&+ H(\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))])] + b,
\end{aligned} \tag{35}$$

where $b \triangleq D_{\mathrm{KL}}(\mathbb{E}[P(\phi(\mathbf{X}_n),\psi(\mathbf{Y}_n))]\|\,\mathcal{O}) - H(P(\phi(\mathbf{X})\mid\psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$, we have:

$$\sup R(\theta^{n\psi}) \geq \sup R(\theta^{\psi}). \tag{36}$$

Proof of incorporating conditions leads to lower generalization risk on learning invariant representations. For the risks of the model having parameters θ^c trained with using conditions, we have:

$$\sup R(\theta^c) = \sup \min_{\phi} [\int L(\theta) dP(\theta) + H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))])] + b, \quad (37)$$

where $b \triangleq D_{\mathrm{KL}}(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))] \| \mathcal{O}) - H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$. For $R(\theta^{nc})$ that trained without using conditions, it has:

$$\begin{aligned} \sup R(\theta^{nc}) &= \sup \min_{\phi,\psi} [\int L(\theta) dP(\theta) + H(\mathbb{E}[P(\phi(\mathbf{X}_n))])] \\ &\quad + H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) \\ &\quad - H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) + b \quad (38) \\ &= \sup \min_{\phi,\psi} [\int L(\theta) dP(\theta) + H(\mathbb{E}[P(\phi(\mathbf{X}_n))])] \\ &\quad + H(\mathbb{E}[P(\psi(\mathbf{Y}_n) \mid \phi(\mathbf{X}_n))]) + b. \end{aligned}$$

Due to the inequality:

$$\begin{aligned} &\sup \min_{\phi} [\, H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]) \,] \quad (39) \\ =& \sup \min_{\phi} [\, H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) \\ &\underbrace{+ H(\mathbb{E}[P(\psi(\mathbf{Y}_n) \mid \phi(\mathbf{X}_n))])}_{\text{Minimized.}} \,] \quad (40) \\ \leq& \sup \min_{\phi,\psi} [\, H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) \,] \\ &\underbrace{+ H(\mathbb{E}[P(\psi(\mathbf{Y}_n) \mid \phi(\mathbf{X}_n))])}_{\text{Remains.}}, \quad (41) \end{aligned}$$

we have

$$\sup R(\theta^{nc}) \geq \sup R(\theta^c). \quad (42)$$

8. Objective derivation details of many previous methods.

This section shows how we uniformly simplify the objectives of previous methods.

ERM [14]: The basic method. The basic method does not focus on minimizing GJSD. Therefore, there are no terms for **Aim 1**. For **Aim 2** it directly minimize $H(P(\phi(\mathbf{X}), \mathbf{Y}))$.

DANN [13]: Minimize feature divergences of source domains. DANN [13] minimizes feature divergences of source domains adverbially without considering conditions. Therefore its empirical objective for **Aim 1** is

$$\min_{\phi} H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) - a \quad (43)$$

For **Aim 2** it directly minimizes $H(P(\phi(\mathbf{X}), \mathbf{Y}))$.

CORAL [48]: Minimize the distance between the second-order statistics of source domains. Since CORAL [48] only minimizes the second-order distance between souce feature distributions, its objective can be summarized as:

$$\min_{\phi} H(P(\phi(\mathbf{X}), \mathbf{Y})) + H(P(\phi(\mathbf{X}))) - H(\mathbb{E}[P(\phi(\mathbf{X}_n))]). \quad (44)$$

By grouping it, CORAL [48] has $-H(\mathbb{E}[P(\phi(\mathbf{X}_n))])$ for **Aim 1** and $H(P(\phi(\mathbf{X}), \mathbf{Y})) + H(P(\phi(\mathbf{X})))$ for **Aim 2**.

CIDG [28]: Minimizing the conditioned domain gap. CIDG [28] tries to learn conditional domain invariant features:

$$\min_{\phi} H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)]). \quad (45)$$

For **Aim 2** it directly minimizes $H(P(\phi(\mathbf{X}), \mathbf{Y}))$.

MDA [16]: Minimizing domain gap compared to the decision gap. Some previous works, such as MDA [16], follow the hypothesis that the generalization is guaranteed while the decision gap is larger than the domain gap. Therefore, instead of directly minimizing the domain gap, MDA minimizes the ratio between the domain gap and the decision gap. The overall objective of MDA can be summarized as:

$$\begin{aligned} &\min_{\phi} H(P(\phi(\mathbf{X}), \mathbf{Y})) + H(P(\phi(\mathbf{X}))) \\ &+ (H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)]) - \underbrace{\mathbb{E}[H(P(\phi(\mathbf{X}_n), \mathbf{Y}_n))]}_{constant}) \\ &- (H(\mathbb{E}[P(\phi(\mathbf{X}_n) \mid \mathbf{Y}_n)]) - \underbrace{\mathbb{E}[H(P(\phi(\mathbf{X}) \mid \mathbf{Y}))]}_{constant}) \\ &+ \underbrace{\mathbb{E}[H(P(\phi(\mathbf{X}), \mathbf{Y}))]}_{constant}. \end{aligned} \quad (46)$$

Since the entropy is non-negative and the constants can be omitted, Eq. 46 is equivalent to:

$$\begin{aligned} &\min_{\phi} H(P(\phi(\mathbf{X}) \mid \mathbf{Y})) + H(P(\phi(\mathbf{X}))) \\ &+ H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)]) - H(\mathbb{E}[P(\phi(\mathbf{X}) \mid \mathbf{Y})]) + a. \end{aligned} \quad (47)$$

By grouping Eq. 47, we have that for **Aim 1** it minimizes $\min_{\phi} H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)])$, and for **Aim 2** it minimizes $H(P(\phi(\mathbf{X}) \mid \mathbf{Y})) - H(\mathbb{E}[P(\phi(\mathbf{X}) \mid \mathbf{Y})])) + H(P(\phi(\mathbf{X})))$.

MIRO [19], SIMPLE [30]: Using pre-trained models as $\mathcal{O}$. One feasible way to obtain $\mathcal{O}$ is adopting pre-trained oracle models such as MIRO [19] and SIMPLE [30]. Note that the pre-trained models are exposed to additional data

besides those provided. Therefore, for **Aim 1**: they have:

$$\min_{\phi} D_{\mathrm{KL}}(P(\phi(X)|Y)\|\mathcal{O}) - \underbrace{\mathbb{E}[H(P(\phi(\mathbf{X}_n), \mathbf{Y}_n))]}_{constant}. \tag{48}$$

Differently, MIRO only uses one pre-trained model, as its $\mathcal{O} \triangleq \mathcal{O}^1$; meanwhile, SIMPLE combines K pre-trained models as the oracle model: $\mathcal{O} \triangleq \sum_{k=1}^{K} v_k \mathcal{O}^k$ where v is the weight vector. For **Aim 2** it directly minimizes $H(P(\phi(\mathbf{X}), \mathbf{Y}))$.

RobustNet [10]. RobustNet employs the instance selective whitening loss, which disentangles domain-specific and domain-invariant properties from higher-order statistics of the feature representation and selectively suppresses domain-specific ones. Therefore, it implicitly whitens the $\mathbf{Y}$-irrelevant features in $\mathbf{X}$. Thus, its objective can be simplified as:

$$\begin{aligned}&\min_{\phi,\psi} H(P(\phi(\mathbf{X}), \psi(\mathbf{Y}))) - H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y})))\\ &+ H(P(\phi(\mathbf{X}))).\end{aligned} \tag{49}$$

9. Aligning notations between paper and supplementary materials

9.1. More details about Table 1

To better understand, we simplify some notations in Table 1. We present the simplified notations and their corresponding origins in Table 9.

9.2. More details about Table 4

For simplification, we uniformly simplified the formulation of terms from their derivation. The simplified form in Table 4 and its original form can be seen in Table 10. Note that the **iAim1** is from CDANN [29], CIDG [28], MDA [16] and **iReg2** is from CORAL [48].

10. Experimental details and parameters

We have conducted 248 **experiments in total,** including 12 Toy experiments (training 3 objective settings on 4 domain settings), 20 Regression experiments in Monocular depth estimation (training 5 objective settings on 4 domain settings), 9 Segmentation experiments (training 3 objective settings on 1 domain settings and verifying on 3 domain settings), 63 Classification experiments (training 1 objective settings on 5 datasets that has $4, 4, 4, 4, 5$ domain settings for 3 trails), and 144 ablation study experiments (training 12 objective settings on 1 dataset that has 4 domain settings for 3 trails). We believe that the consistent improvements yielded by GMDG in these experiments validate the superiority of our GMDG.

Experimental details of these experiments can be found in the following. Note that we set $v_{A2} = 1$ for all experiments.

10.1. Toy experiments: Synthetic regression experimental details.

We explore the efficacy of ψ by using toy regression experiments with synthetic data.

Datasets. The latent features in all three domains are added some distributional shifts and used as the first group in the raw features (denoted as x_n^1, y_n^1 where $n \in 1, 2, 3$ represent which domain it belongs to). Then, some domain-conditioned transformations are applied to shifted features, or some pure random noises are used as the second group in the raw features (denoted as x_n^2, y_n^2). Therefore the constructed $X_{n\in\{1,2,3\}} = [x_n^1, x_n^2], Y_{n\in\{1,2,3\}} = [y_n^1, y_n^2]$ both contain features that dependents on D. Details of each synthetic data are exhibited in Table 11. We generate 10000 samples for training and 100 samples for validation and testing sets.

Parameter settings. All experiments are conducted with $v_{A1}, v_{R1}, v_{R2} = 0.1$.

Experimental settings. For ϕ, ψ, we use three-layer MLP and one linear layer for regression prediction and Mean Squared Error (MSE) as the loss. We use the best model on the validation dataset for testing.

Metric. We use the MSE between the predictions and the Y of the testing set as the evaluation metric.

10.2. Regression experiments: Monocular depth estimation details.

We explore the efficacy of ψ with GMDG by using toy monocular depth estimation experiments with NYU Depth V2 dataset [46].

Datasets. NYU Depth V2 contains images with 480×640 resolution with depth values ranging from 0 to 10 meters. We adopt the densely labeled pairs for training and testing.

Multi-domain construction. To construct multiple domains that fit the problem settings, we split the NYU Depth V2 dataset into four categories as four domains:

- School: study room, study, student lounge, printer room, computer lab, classroom.
- Office: reception room, office kitchen, office, nyu office, conference room.
- Home: playroom, living room, laundry room, kitchen, indoor balcony, home storage, home office, foyer, dining room, dinette, bookstore, bedroom, bathroom, basement.
- Commercial:furniture store, exercise room, cafe.

After filtering data samples that are not able to be used, each domain has 95, 110, 1209, and 35 data pairs that can be used for training.

	Learning domain invariant representations Aim1: Learning domain invariance	Reg1: Integrating prior
DANN	$\min_\phi H(P(\phi(\mathbf{X}) \mid \mathcal{D}))$ $\min_\phi H(\mathbb{E}[P(\phi(\mathbf{X}_n))]$	None
CDANN, CIDG, MDA	$\min_\phi H(P(\phi(\mathbf{X}), \mathbf{Y} \mid \mathcal{D}))$ $\min_\phi H(\mathbb{E}[P(\phi(\mathbf{X}_n), \psi(\mathbf{Y}_n))]$	None
Ours	$\min_{\phi,\psi} H(P(\phi(\mathbf{X}), \psi(\mathbf{Y}) \mid \mathcal{D}))$ $\min_{\phi,\psi} H(\mathbb{E}[P(\phi(\mathbf{X}_N), \psi(\mathbf{Y}_n)])$	$\min_{\phi,\psi} D_{\mathrm{KL}}(P(\phi(\mathbf{X}), \psi(\mathbf{Y}))\|\mathcal{O})$

	Maximizing A Posterior between representations and targets Aim2: Maximizing A Posterior (MAP)	Reg2: Suppressing invalid causality
CORAL	$\min_\phi H(P(\mathbf{Y} \mid \phi(\mathbf{X})))$	$\min_\phi -H(P(\phi(\mathbf{X} \mid \mathcal{D}))) + H(P(\phi(\mathbf{X})))$ $\min_\phi -H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) + H(P(\phi(\mathbf{X})))$

Table 9. Supplemental notations for Table 1. Refined notations and their original formulations are reported. The original formulations are highlighted as blue.

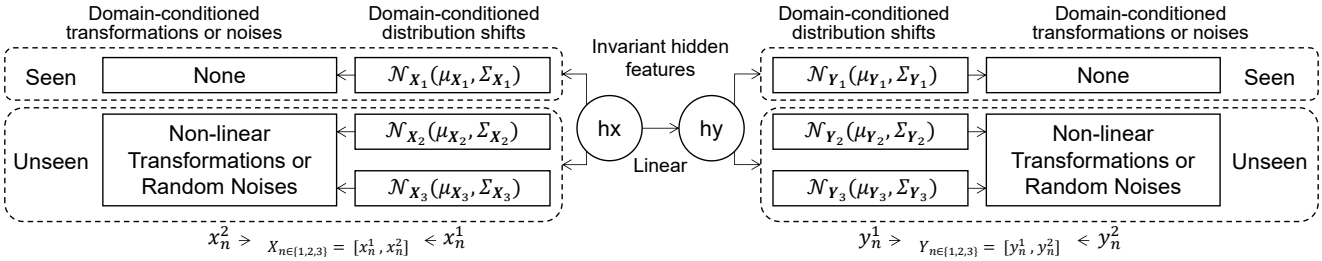

Figure 3. Toy experiments: Diagram of constructing the toy dataset.

GAim2	$H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X}))) + H(P(\mathbf{Y} \mid \psi(\mathbf{Y})))$
GReg1	$D_{\mathrm{KL}}(P(\phi(\mathbf{X}), \mathbf{Y} \mid \mathcal{D})\|\mathcal{O}))$
iAim1	$H(P(\phi(\mathbf{X}) \mid \mathcal{D}))$
GAim1	$H(P(\phi(\mathbf{X}), \mathbf{Y}) \mid \mathcal{D})$
iReg2	$-H(P(\phi(\mathbf{X}), \mathcal{D}) + H(P(\phi(\mathbf{X})))$
GReg2	$-H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$
GAim2	$H(P(\psi(\mathbf{Y}) \mid \phi(\mathbf{X}))) + H(P(\mathbf{Y} \mid \psi(\mathbf{Y})))$
GReg1	$D_{\mathrm{KL}}(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)]\|\mathcal{O}))$
iAim1	$H(\mathbb{E}[P(\phi(\mathbf{X}_n))])$
GAim1	$H(\mathbb{E}[P(\phi(\mathbf{X}_n), \mathbf{Y}_n)])$
iReg2	$-H(\mathbb{E}[P(\phi(\mathbf{X}_n))]) + H(P(\phi(\mathbf{X})))$
GReg2	$-H(P(\phi(\mathbf{X}) \mid \psi(\mathbf{Y}))) + H(P(\phi(\mathbf{X})))$

Table 10. Notations for terms in the paper (above) and its derived formulation (below) in the appendix.

Parameter settings. We follow all the hyperparameter settings in the VA-DepthNet and set $v_{A1} = 0.001, v_{R1} = 0.001, v_{R2} = 0.0001$. Note that the backbone is trained using VA-DepthNet but without the Variational Loss proposed by VA-DepthNet.

Experimental settings. We use the final saved checkpoint for the leave-one-out cross-validation.

Metrics. Please see metric details in VA-DepthNet [32].

10.3. Segmentation experimental details.

We follow the experimental settings of RobustNet for segmentation experiments.

Datasets. There are two groups of datasets: Synthetic datasets and real-world datasets. (1) Synthetic datasets: GTAV [41] is a large-scale dataset containing 24,966 driving-scene images generated from the Grand Theft Auto V game engine. SYNTHIA [42] which is composed of photo-realistic synthetic images has 9,400 samples with a resolution of 960×720. (2) Real-world datasets: Cityscapes [11] is a large-scale dataset containing high-resolution urban scene images. Providing 3,450 finely annotated images and 20,000 coarsely annotated images, it collects data from 50 different cities in primarily Germany. Only the finely annotated set is adopted for our training and validation. BDD-100K [53] is another real-world dataset

hx	$hx \sim \mathcal{N}(hx; 0, 1)$	
hy	$hy = hx$	
Data 1	Without distribution shift	With affine transformations
X_1	$x_1^1 = hx$	$x_1^2 = x_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
Y_1	$y_1^1 = hy$	$y_1^2 = y_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
X_2	$x_2^1 = hx$	$x_2^2 = 4 \times x_2^1 + \epsilon \sim \mathcal{N}(\epsilon; 0.5, 0.3)$
Y_2	$y_2^1 = hy$	$y_2^2 = 4 \times y_2^1 + 0.3$
X_3	$x_3^1 = hx$	$x_3^2 = 2 \times x_3^1 + \epsilon \sim \mathcal{N}(\epsilon; -0.3, 0.2)$
Y_3	$y_3^1 = hy$	$y_3^2 = 0.5 \times y_3^1 - 0.2$
Data 2	With distribution shift	With affine transformations
X_1	$x_1^1 = hx$	$x_1^2 = x_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
Y_1	$y_1^1 = hy$	$y_1^2 = y_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
X_2	$x_2^1 = hx + \epsilon \sim \mathcal{N}(\epsilon; -0.1, 0.1)$	$x_2^2 = 4 \times x_2^1 + \epsilon \sim \mathcal{N}(\epsilon; 0.3, 0.3)$
Y_2	$y_2^1 = hy + \epsilon \sim \mathcal{N}(\epsilon; 0.2, 0.1)$	$y_2^2 = 8 \times y_2^1 - 0.3$
X_3	$x_3^1 = hx + \epsilon \sim \mathcal{N}(\epsilon; 0.4, 0.2)$	$x_3^2 = -1 \times x_3^1 + \epsilon \sim \mathcal{N}(\epsilon; -0.3, 0.2)$
Y_3	$y_3^1 = hy + \epsilon \sim \mathcal{N}(\epsilon; -0.4, 0.2)$	$y_3^2 = \epsilon \sim \mathcal{N}(\epsilon; 0, 0.2)$
Data 3	Without distribution shift	With squared, cubed transformations or noises
X_1	$x_1^1 = hx$	$x_1^2 = x_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
Y_1	$y_1^1 = hy$	$y_1^2 = y_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
X_2	$x_2^1 = hx$	$x_2^2 = 4 \times x_2^1 ** 3 + \epsilon \sim \mathcal{N}(\epsilon; 0.5, 0.3)$
Y_2	$y_2^1 = hy$	$y_2^2 = 4 \times y_2^1 ** 2 + 0.3$
X_3	$x_3^1 = hx$	$x_3^2 = 2 \times x_3^1 ** 2 + \epsilon \sim \mathcal{N}(\epsilon; -0.3, 0.2)$
Y_3	$y_3^1 = hy$	$y_3^2 = 0.5 \times y_3^1 ** 3 - 0.2$
Data 4	With distribution shift	With squared, cubed transformations or noises
X_1	$x_1^1 = hx$	$x_1^2 = x_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
Y_1	$y_1^1 = hy$	$y_1^2 = y_1^1 + \epsilon \sim \mathcal{N}(\epsilon; 0, 0.3)$
X_2	$x_2^1 = hx + \epsilon \sim \mathcal{N}(\epsilon; -0.1, 0.1)$	$x_2^2 = 4 \times x_2^1 ** 3 + \epsilon \sim \mathcal{N}(\epsilon; 0.5, 0.3)$
Y_2	$y_2^1 = hy + \epsilon \sim \mathcal{N}(\epsilon; 0.2, 0.1)$	$y_2^2 = 4 \times y_2^1 ** 2 + 0.3$
X_3	$x_3^1 = hx + \epsilon \sim \mathcal{N}(\epsilon; 0.4, 0.2)$	$x_3^2 = 2 \times x_3^1 ** 2 + \epsilon \sim \mathcal{N}(\epsilon; -0.3, 0.2)$
Y_3	$y_3^1 = hy + \epsilon \sim \mathcal{N}(\epsilon; -0.4, 0.2)$	$y_3^2 = 0.5 \times y_3^1 ** 3 - 0.2$

Table 11. Toy experiments: Synthetic data details for each experiment.

with a resolution of 1280×720. It provides diverse urban driving scene images from various locations in the US. We use the 7,000 training and 1,000 validation of the semantic segmentation task. The images are collected from various locations in the US. Mapillary is also a real-world dataset that contains worldwide street-view scenes with 25,000 high-resolution images.

Parameter settings. Specifically, we use all RobustNet's hyper-parameters and set $v_{A1} = 0.0001, v_{R1} = 0.0001$.

10.4. Classification experimental details.

Datasets. We use PACS (4 domains, 9,991 samples, 7 classes) [25], VLCS (4 domains, 10,729 samples, 5 classes) [12], OfficeHome (4 domains, 15,588 samples, 65 classes) [50], TerraIncognita (TerraInc, 4 domains, 24,778 samples, 10 classes) [2], and DomainNet (6 domains, 586,575 samples, 345 classes) [38].

Parameter settings. We list the hyper-parameters in Table 12 to reproduce our results.

Metric. We employ mean Intersection over Union (mIoU) as the measurement for the segmentation task.

10.5. Ablation studies experimental details.

Parameter settings. We run each experiment in three trials with seeds: $[0, 1, 2]$. Full settings are reported in Table 13. Especially,

Experimental settings. We use SWAD for all ablation studies to alleviate the effeteness of hyper-parameters. All ablation studies share the same hyper-parameters but add different combinations of terms. CORAL's [48] objective focuses on minimizing the learned feature covariance discrepancy between source and target, requiring target data access and only regards second-order statistics. We adapt

Use ResNet-50 without SWAD	$v2$	$v3$	$v1$	lr mult	lr	dropout	WD	TR	CF
TerraIncognita	0.1	0.1	0.2	12.5	-	-	-	-	-
OfficeHome	0.1	0.001	0.1	20.0	3e-5	0.1	1e-6	-	-
VLCS	0.01	0.001	0.1	10.0	1e-5	-	1e-6	0.2	50
PACS	0.01	0.01	0.01	25.0	-	-	-	-	-
DomainNet	0.1	0.1	0.1	7.5	-	-	-	-	500

Use ResNet-50 with SWAD	$v2$	$v3$	$v1$	lr mult	CF
TerraIncognita	0.1	0.001	0.01	10.0	-
OfficeHome	0.1	0.1	0.3	10.0	-
VLCS	0.01	0.001	0.1	10.0	50
PACS	0.01	0.001	0.1	20.0	-
DomainNet	0.1	0.1	0.1	7.5	500

Use RegNetY-16GF with and without SWAD	$v2$	$v3$	$v1$	lr mult	CF
TerraIncognita	0.01	0.01	0.01	2.5	-
OfficeHome	0.01	0.1	0.1	0.1	-
VLCS	0.01	0.01	0.1	2.0	50
PACS	0.01	0.1	0.1	0.1	-
DomainNet	0.1	0.1	0.1	7.5	500

Table 12. Classification experiments: Parameter settings of classification tasks. Notations: WD: weight decay; TR: tolerance ratio; CF: checkpoint freq. - denotes that for where the default settings are used.

Ablation studies on OfficeHome	$v2$	$v3$	$v1$	lr mult	use **iAim1**	use **iReg2**
Base (ERM)	0.0	0.0	0.0	0.1	False	False
Base +iAim1 (DANN)	0.0	0.0	0.1	0.1	True	False
Base + GAim1 (CDANN, CIDG)	0.0	0.0	0.1	0.1	False	False
Base +iReg2 (CORAL+ψ)	0.0	0.1	0.0	0.1	False	True
Base + GReg2	0.0	0.1	0.0	0.1	False	False
Base + GAim1 + GReg2 (MDA+ψ)	0.0	0.1	0.1	0.1	False	False
Base + GReg1 (MIRO, SIMPLE)	0.01	0.0	0.0	0.1	False	False
Base + GReg1 +iAim1	0.01	0.0	0.1	0.1	False	True
Base + GReg1 + GAim1	0.01	0.0	0.1	0.1	False	False
Base + GReg1 +iReg2	0.01	0.1	0.0	0.1	True	False
Base + GReg1 + GReg2	0.01	0.1	0.0	0.1	False	False
Base + GReg1 + GAim1 + GReg2 (Ours)	0.01	0.1	0.1	0.1	False	False

Table 13. Ablation studies: Parameter settings of ablation studies. Notations: WD: CF: checkpoint freq. - denotes that for where the default settings are used.

its approach to minimize feature covariances across seen domains for a fair comparison.

11. More results

Visualization of toy experiments: Please see the visualization of toy experiments in Figure 4.

Regression results: Monocular depth estimation. The regression results for each unseen domain of monocular depth estimation visualization is displayed in Figure 6, 7.

The Visualization of regression results for unseen domains of models trained with different objective settings are exhibited in Figure 8, 9.

Segmentation results. The segmentation results for unseen samples are displayed in Figure 10.

Classification results. We show the results of each category for the classification experiments as Table 15, 16, 17, 18, 19.

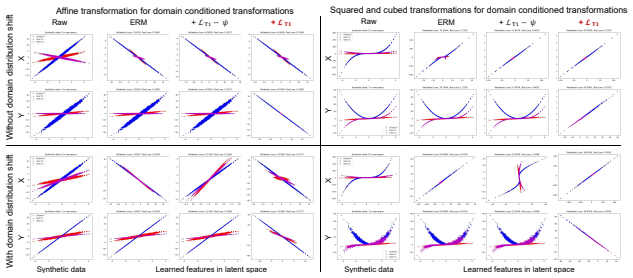

Figure 4. Toy experiments: Visualization of learned latent representations of different methods. Each color represents a domain.

11.1. Other findings and Analysis

What makes a better $\mathcal{O}$. As demonstrated in Eq. 7, $\mathcal{O}$ plays a crucial role in PUB by anchoring a space where the relationship between $\mathbf{X}$ and $\mathbf{Y}$ is preserved. Ideally, having one $\mathcal{O}$ that provides general representations for all seen and unseen domains leads to the best results, one finding supported by MIRO and SIMPLE. However, even though SIMPLE++ combines 283 pre-trained models, achieving the 'perfect' $\mathcal{O}$ remains unattained. Therefore, this paper primarily discusses how our proposed objectives can improve the model performance when a fixed $\mathcal{O}$ is provided.

Comparison with MDA: Minimizing domain gap compared to the decision gap. MDA [16], guided by the hypothesis "guaranteed generalization only when the decision gap exceeds the domain gap", aims to minimize the ratio between the domain gap and the decision gap. This approach facilitates learning $\mathcal{D}$-independent conditional features, enhancing class separability across domains. As Table 1 illustrates, MDA's **Reg2** objective can also be interpreted as suppressing invalid causality, aligning with our approach. However, MDA's implementation requires manual selection of $\phi(\mathbf{X})$ from the same $\mathbf{Y}$ without using ψ and **GReg2**. Our method further relaxes MDA's assumption, extending the application of the objective and making it also applicable to tasks besides classification, such as segmentation.

Cutting off causality form $\phi(\mathbf{X}) \rightarrow \psi(\mathbf{Y})$ may lead to collapse of the model. We have tried to reversely suppress the causality form $\phi(\mathbf{X}) \rightarrow \psi(\mathbf{Y})$ instead of causality form $\psi(\mathbf{Y}) \rightarrow \phi(\mathbf{X})$ for monocular depth estimation in VA-DepthNet and it causes collapse.

Suppressing invalid causality: Why this design: Our GMDG introduces a mapping ψ for $\mathbf{Y}$ to relax the static assumption, corroborating more general and practical scenarios. Our empirical findings, as shown in Fig. 5, reveal that introducing $\psi(\mathbf{Y})$ without any constraints may not guarantee a clear decision margin for classification. Upon examining our objective in Eq.8 in the main manuscript, we hypothesize that the effect might result from '$\psi(\mathbf{Y})$ causing $\phi(\mathbf{X})$', which we term 'invalid causality'. Thus, we designed a term to suppress such invalid causality. This term is derived from the prediction perspective wherein $\mathbf{Y}$ should be only predicted from $\phi(\mathbf{X})$; hence, $\phi(\mathbf{X})$ should not be caused by $\psi(\mathbf{Y})$. Consider the scenario wherein $\mathbf{Y}$ and $\psi(\mathbf{Y})$ are absent during prediction - the hypothesized causality from $\psi(\mathbf{Y})$ to $\phi(\mathbf{X})$ would disrupt the causal chain, resulting in an 'incomplete' representation of $\phi(\mathbf{X})$ then prediction degradation. Hence, it is critical to suppress $\psi(\mathbf{Y}) \not\rightarrow \phi(\mathbf{X})$ that may occur during joint training. Notably, the suppression is not symmetric and promotes $\phi(\mathbf{X}) \rightarrow \mathbf{Y}$. **Intuition:** Intuitively, **GReg2** further 'erases' the redundant information in $\phi(\mathbf{X})$ that may be caused by $\phi(\mathbf{Y})$, which aims to refine the latent space, yielding better invariant latent features for predicting $\mathbf{Y}$ (i.e., larger decision margin for the unseen domain as highlighted in Fig. 5).

GMDG's efficiency: We have analyzed the efficiency of GMDG in Tab. 14. Though theoretically superior in generality, GMGD may increase computational costs during training due to additional loss functions and VAE encoders. However, these auxiliary components are discarded in the inference stage, ensuring their efficiency remains unaffected. Our confidence in the model's amenability to efficiency enhancement through careful design is high, and such pursuits remain a promising avenue for future work.

Applicability, constraint, and limitations: GMDG is specifically designed for mDG with accessible $P(\mathbf{Y})$, in which the model is trained on multiple seen domains and tested on one unseen domain. This task essentially requires learning the invariance across multi-domains for the prediction. When used in single-domain generalization or cases involving novel classes in the unseen domain, GMDG may not be directly applicable, thus requiring further investigations. Meanwhile, as a general objective, our novel GMDG involves additional modules/losses that may incur extra computational costs during training. We have discussed these aspects in our main paper, and we would like to leave them as future work.

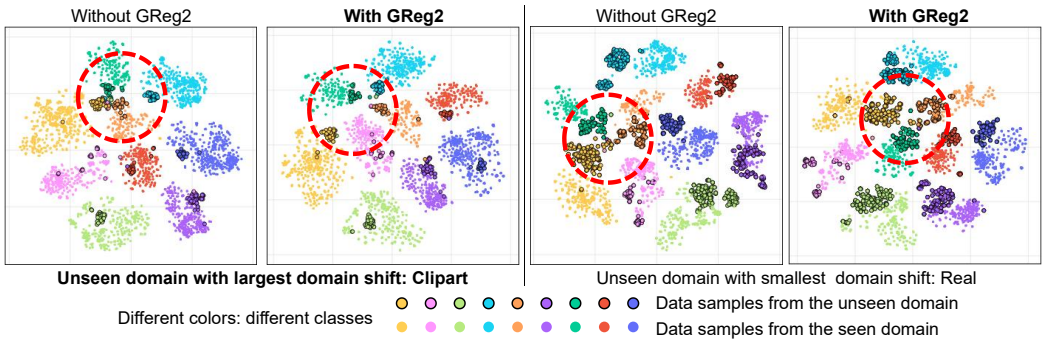

Figure 5. T-SNE map of latent features from classification models that were trained without and with **GReg2**.

	Classification		Depth estimation		Segmentation		
	Training	Inference	Training	Inference	Training	Inference	
FLOPs (G)	99.16	49.58	584.44	573.78	209.97	167.34	Baseline
	124.00	49.58	1543.90	573.78	449.83	167.34	With GMDG
Parameters (M)	2.94	1.47	64.42	64.27	23.13	22.54	Baseline
	4.93	1.47	123.76	64.27	82.48	22.54	With GMDG

Table 14. FLOPs and parameters of baselines without and with GMDG during training and inference.

TerraIncognita	**Location 100**	**Location 38**	**Location 43**	**Location 46**	**Avg.**
ERM [14]	54.3	42.5	55.6	38.8	47.8
MIRO [19] (use ResNet-50)	-	-	-	-	50.4
GMDG (use ResNet-50)	59.8±1.0	45.3±1.7	57.1±1.8	38.2±5	50.1±1.2
ERM + SWAD [6]	55.4	44.9	59.7	39.9	50.0
DIWA [40]	57.2	50.1	60.3	39.8	51.9
MIRO [19] + SWAD [6] (use ResNet-50)	-	-	-	-	52.9
GMDG + SWAD (use ResNet-50)	61.2±1.4	48.4±1.6	60.0±0.4	42.5±1.1	53.0±0.7
MIRO [19] (use RegNetY-16GF)	-	-	-	-	58.9
GMDG (use RegNetY-16GF)	73.3±3.3	54.7±1.4	67.1±0.3	48.6±6.5	60.7±1.8
MIRO [19] + SWAD [6] (use RegNetY-16GF)	-	-	-	-	64.3
GMDG + SWAD (use RegNetY-16GF)	**74.3±1.5**	**59.2±1.2**	**70.6±1.1**	**56.0±0.8**	**65.0±0.2**

Table 15. Classification experiments on TerraIncognita: More results of full GMDG for each category.

OfficeHome	**art**	**clipart**	**product**	**real**	**Avg.**
ERM [14]	63.1	51.9	77.2	78.1	67.6
MIRO [19] (use ResNet-50)	-	-	-	-	70.5±0.4
GMDG (use ResNet-50)	68.9±0.3	56.2±1.7	79.9±0.6	82.0±0.4	70.7±0.2
ERM + SWAD [6]	66.1	57.7	78.4	80.2	70.6
DIWA [40]	69.2	59	81.7	82.2	72.8
MIRO [19] + SWAD [6] (use ResNet-50)	-	-	-	-	72.4±0.1
GMDG + SWAD (use ResNet-50)	68.9±0.6	58.2±0.6	80.4±0.3	82.6±0.4	72.5±0.2
MIRO [19] (use RegNetY-16GF)	-	-	-	-	80.4±0.2
GMDG (use RegNetY-16GF)	79.7±1.6	67.7±1.8	87.8±0.8	87.9±0.7	80.8±0.6
MIRO [19] + SWAD [6] (use RegNetY-16GF)	-	-	-	-	83.3±0.1
GMDG + SWAD (use RegNetY-16GF)	**84.1±0.2**	**74.3±0.9**	**89.9±0.4**	**90.6±0.1**	**84.7±0.2**

Table 16. Classification experiments on OfficeHome: More results of full GMDG for each category.

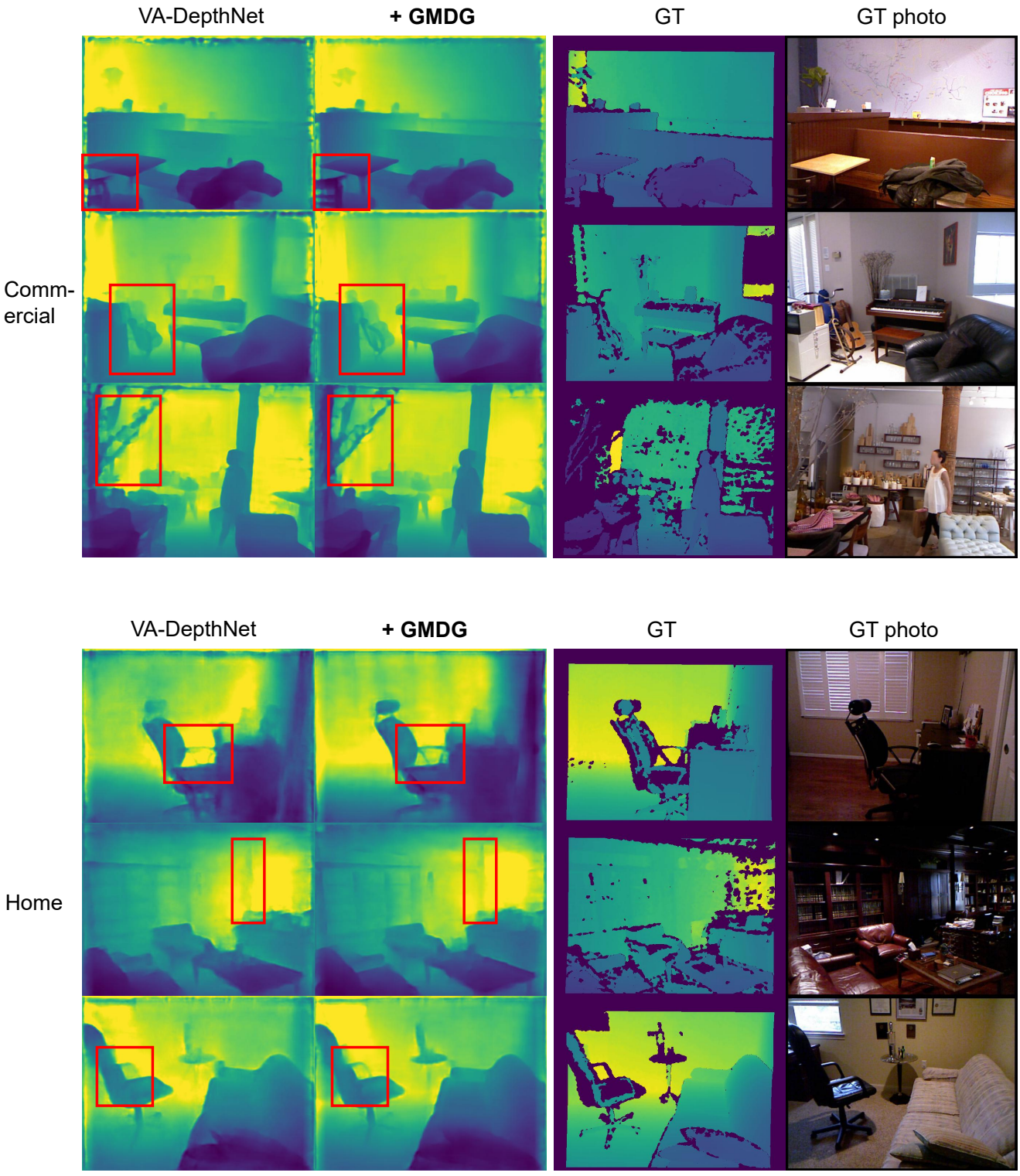

Figure 6. Regression results: Monocular depth estimation results between VA-Depth and our GMDG on samples from unseen domains. It can be seen that better generalization across domains is obtained with GMDG.

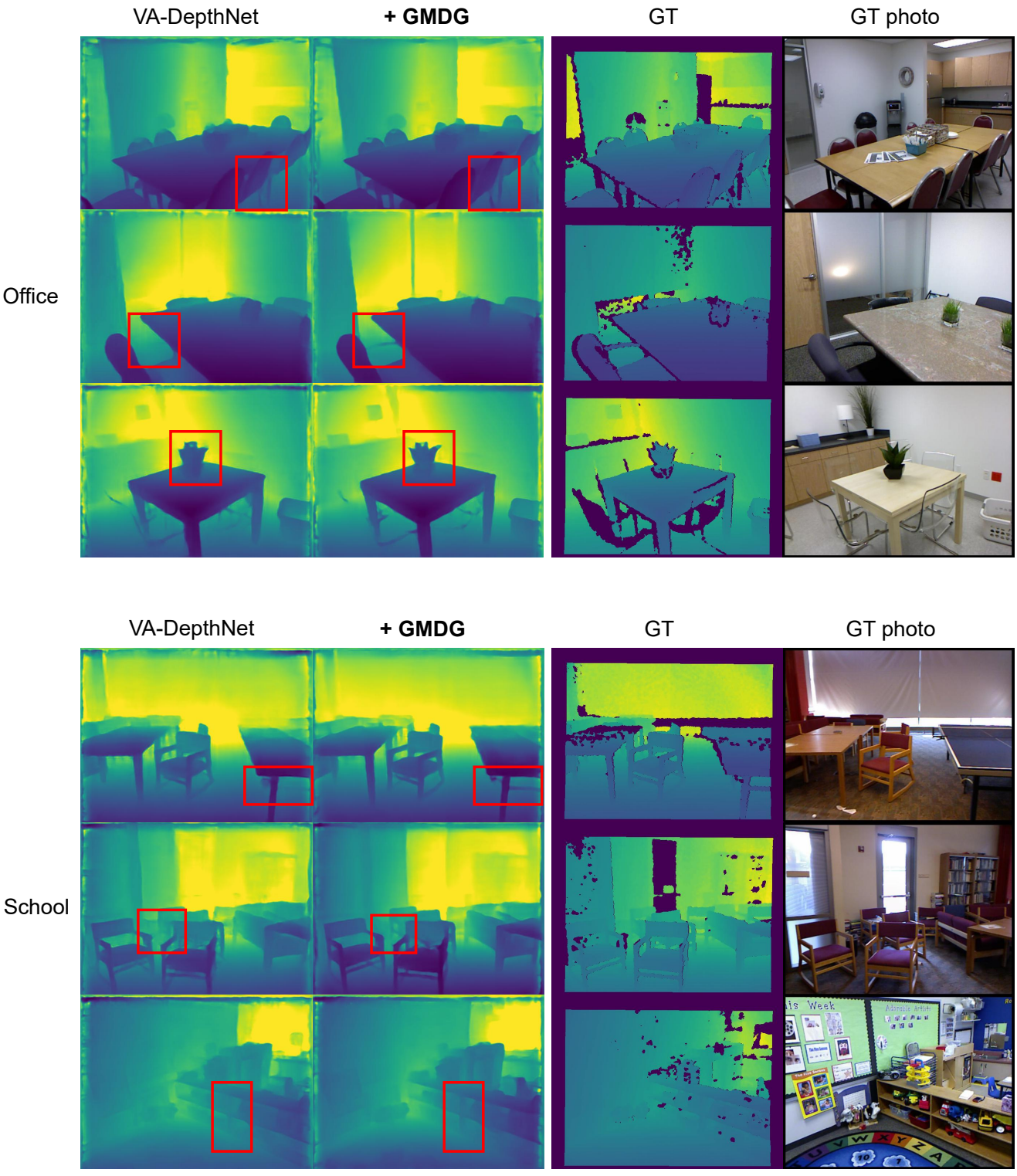

Figure 7. Regression results: Monocular depth estimation results between VA-Depth and our GMDG on samples from unseen domains. It can be seen that better generalization across domains is obtained with GMDG, continues.

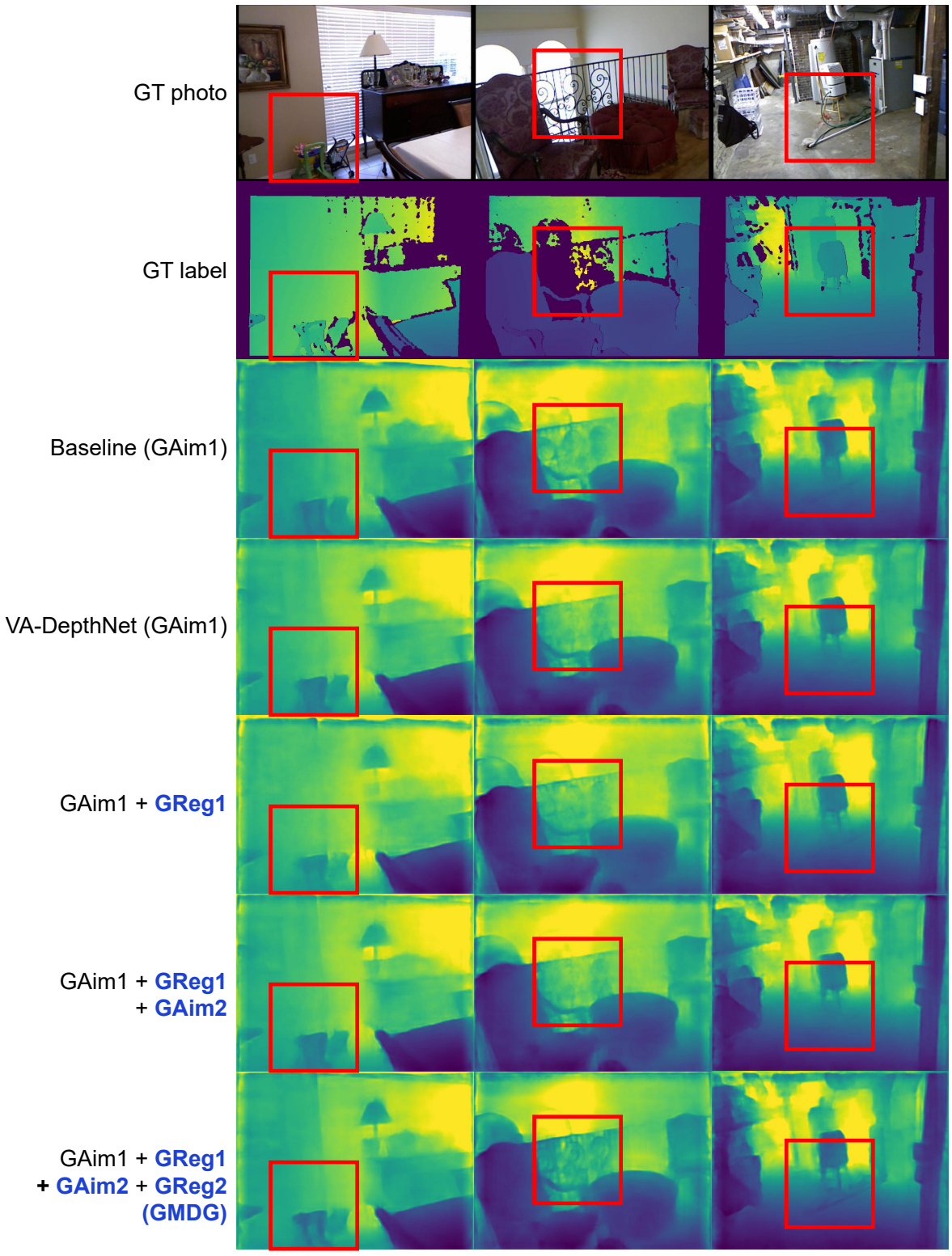

Figure 8. Regression results: Monocular depth estimation results for unseen domains of models trained with different objective settings. It can be seen that with the whole GMDG, the model performs the best generalization for all unseen domain settings.

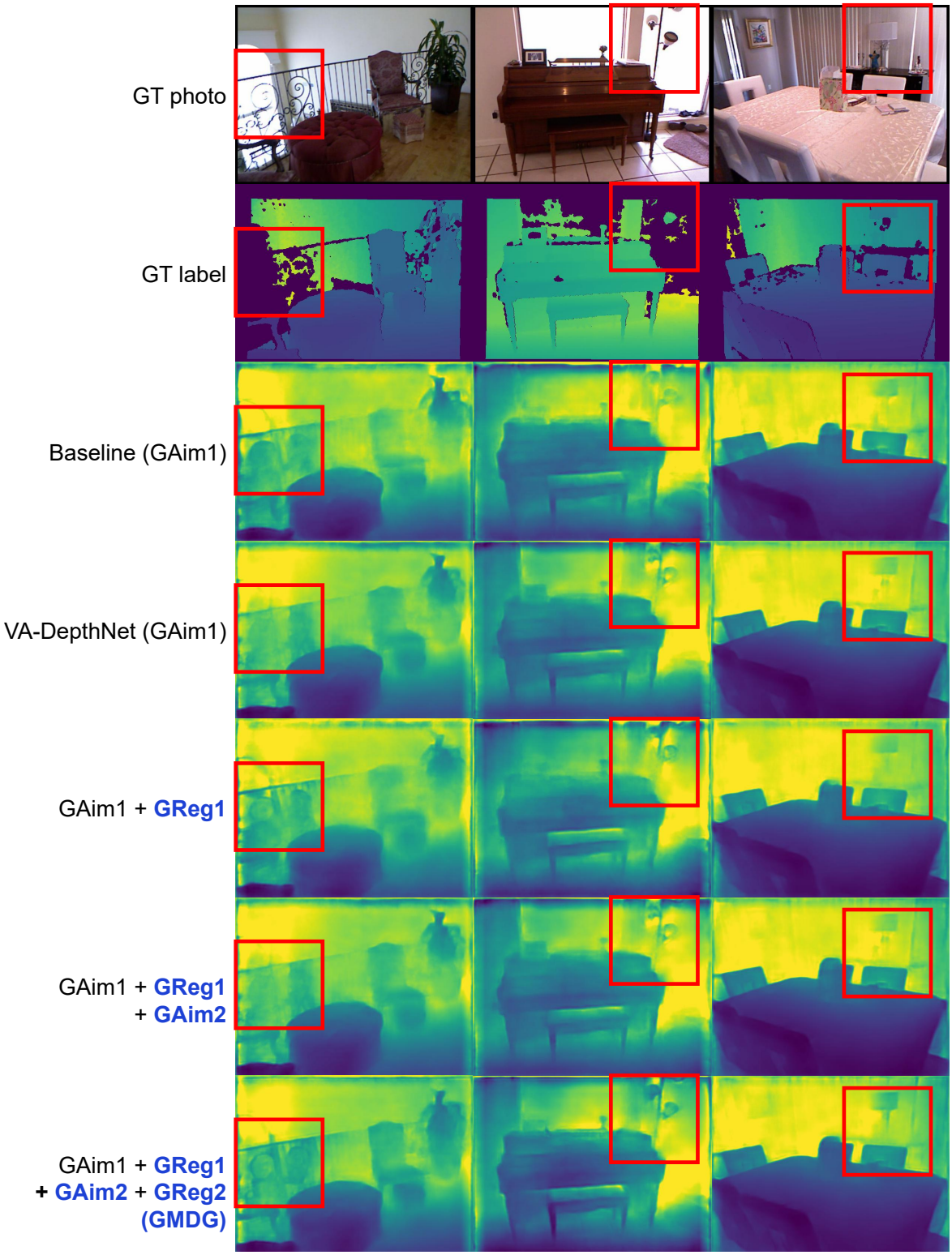

Figure 9. Regression results: Monocular depth estimation results for unseen domains of models trained with different objective settings, continues. It can be seen that with the whole GMDG, the model performs the best generalization for all unseen domain settings.

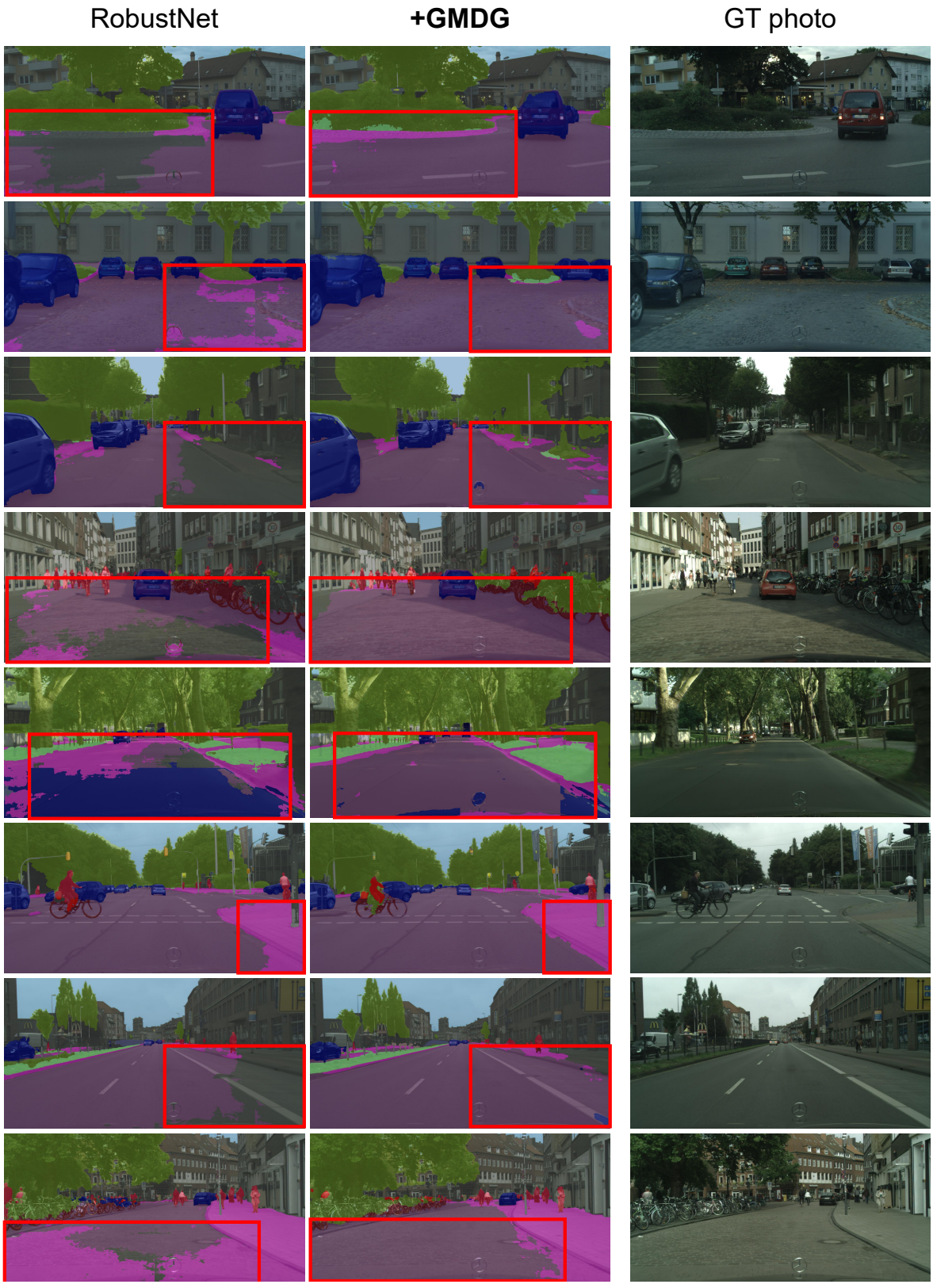

Figure 10. Segmentation results: Visualizations between RobostNet and our GMDG on samples from unseen domains. It can be seen that better generalization is obtained with GMDG.

VLCS	caltech101	labelme	sun09	voc2007	Avg.
ERM [14]	97.7	64.3	73.4	74.6	77.3
MIRO [19] (use ResNet-50)	-	-	-	-	79.0±0.0
GMDG (use ResNet-50)	98.3±0.4	65.9±1	73.4±0.8	79.3±1.3	79.2±0.3
ERM + SWAD [6]	98.8	63.3	75.3	79.2	79.1
DIWA [40]	98.9	62.4	73.9	78.9	78.6
MIRO [19] + SWAD [6](use ResNet-50)	-	-	-	-	79.6±0.2
GMDG + SWAD (use ResNet-50)	98.9±0.4	63.6±0.2	76.4±0.5	79.5±0.6	79.6±0.1
MIRO [19] (use RegNetY-16GF)	-	-	-	-	79.9±0.6
GMDG (use RegNetY-16GF)	97.9±1.3	**66.8±2.1**	**80.8±1**	83.9±1.8	**82.4±0.6**
MIRO [19] + SWAD [6] (use RegNetY-16GF)	-	-	-	-	81.7±0.1
GMDG + SWAD (use RegNetY-16GF)	**98.4±0.1**	65.5±1.4	79.9±0.4	**84.9±0.9**	82.2±0.3

Table 17. Classification experiments on VLCS: More results of full GMDG for each category.

PACS	art_painting	cartoon	photo	sketch	Avg.
ERM [14]	84.7	80.8	97.2	79.3	84.2
MIRO [19] (use ResNet-50)	-	-	-	-	85.4±0.4
GMDG (use ResNet-50)	84.7±1.0	81.7±2.4	97.5±0.4	80.5±1.8	85.6±0.3
ERM + SWAD [6]	89.3	83.4	97.3	82.5	88.1
DIWA [40]	90.6	83.4	98.2	83.8	89.0
MIRO [19] + SWAD [6] (use ResNet-50)	-	-	-	-	88.4±0.1
GMDG + SWAD (use ResNet-50)	90.1±0.6	83.9±0.2	97.6±0.5	82.3±0.7	88.4±0.1
MIRO [19] (use RegNetY-16GF)	-	-	-	-	97.4±0.2
GMDG (use RegNetY-16GF)	97.5±1.0	97.0±0.2	99.4±0.2	95.2±0.4	97.3±0.1
MIRO [19] + SWAD [6] (use RegNetY-16GF)	-	-	-	-	96.8±0.2
GMDG + SWAD (use RegNetY-16GF)	**98.3±0.3**	**98.0±0.1**	**99.5±0.3**	**95.3±0.8**	**97.9±0.0**

Table 18. Classification experiments on PACS: More results of full GMDG for each category.

DomainNet	clipart	info	painting	quickdraw	real	sketch	Avg.
ERM [14]	50.1	63.0	21.2	63.7	13.9	52.9	44.0
MIRO [19] (use ResNet-50)	-	-	-	-	-	-	44.3±0.2
GMDG (use ResNet-50)	63.4±0.3	22.4±0.4	51.4±0.4	13.4±0.8	64.4±0.3	52.4±0.4	44.6±0.1
ERM + SWAD [6]	53.5	66.0	22.4	65.8	16.1	55.5	46.5
DIWA [40]	55.4	66.2	23.3	68.7	16.5	56.0	47.7
MIRO [19] + SWAD [6] (use ResNet-50)	-	-	-	-	-	-	47.0±0.0
GMDG + SWAD (use ResNet-50)	66.4±0.3	23.8±0.1	54.5±0.3	15.8±0.1	67.5±0.1	55.8±0.0	47.3±0.1
MIRO [19] (use RegNetY-16GF)	-	-	-	-	-	-	53.8±0.1
GMDG (use RegNetY-16GF)	74.0±0.3	39.5±1.5	61.5±0.3	16.3±1.2	73.9±1.5	62.8±2.4	54.6±0.1
MIRO [19] + SWAD [6] (use RegNetY-16GF)	-	-	-	-	-	-	60.7±0.0
GMDG + SWAD (use RegNetY-16GF)	**79.0±0.0**	**46.9±0.4**	**69.9±0.4**	**20.7±0.6**	**81.1±0.3**	**70.3±0.4**	**61.3±0.2**

Table 19. Classification experiments on DomainNet: More results of full GMDG for each category.