Dual Masked Generative Adversarial Transformer for Unsupervised Domain Adaptation

Lei Zhu[1*], Xinxing Xu[2], Jun Zhou[3], Qiegen Liu[1], Rick Siow Mong Goh[3], and Yong Liu[3]

[1] School of Information Engineering, Jiangxi Provincial Key Laboratory of Advanced Signal Processing and Intelligent Communications, Nanchang University, China
[2] Microsoft Research Asia, Singapore
[3] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore

Abstract. Transformer has recently garnered significant interest in unsupervised domain adaptation tasks due to its superior generalization ability. State-of-the-art methods leverage masked image modeling with consistency regularization to improve target domain performance. However, such strategy becomes less effective when the domain gap becomes large, as the domain gap between masked domain and original domain is not minimized. How to address the adaptation problem when domain gap becomes large is an important research problem in domain adaptation. In this paper, we propose Dual Masked Generative Adversarial Transformer (**Dual-MGAT**), which simultaneously aligns the masked target and masked source domain towards source and target domain to address the large domain gap problem. Specifically, we formulate a general framework for masked domain learning. To reduce the domain gap between the masked domain (e.g. masked target) and the original domain (e.g. source), we note that there exists two types of discrepancies, namely the information gap due to masked image modeling and the domain distribution discrepancy. To this end, we propose the masked generative adversarial adaptation technique, which introduces a transformer decoder to bridge the masked feature space towards the original feature space via [CLS] token feature generation and utilizes a class conditional domain discriminator for adversarial distribution alignment to reduce the domain distribution discrepancy. We further investigate the dual branch of masked source to target domain adaptation, which boosts performance. We provide a theoretical analysis of our framework from a masked domain adaptation perspective. Extensive experimental studies demonstrate the superiority of our framework.

Keywords: Vision Transformer · Masked Generative Adversarial Adaptation · Domain Adaptation

* Corresponding author. Email: zhulei@ncu.edu.cn

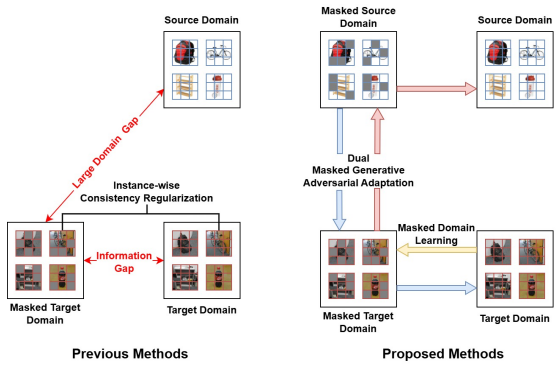

Fig. 1: Conceptual comparison between previous methods and our proposed method in unsupervised domain adaptation with masked image modeling. **Previous methods** leverage masked image modeling for consistency regularization at target data instance level. However, they overlook the domain gap between masked and original domain, which undermines domain adaptation performance. Our **proposed method** constructs and learns masked domains to assist domain adaptation and performs dual masked generative adversarial adaptation to simultaneously align the masked target and masked source domain towards source and target domain to effectively address the large domain gap problem.

1 Introduction

Deep learning has achieved remarkable success in various computer vision tasks; however, it requires large-scale labeled data to be effective, which is both time-consuming and expensive to obtain. Unsupervised domain adaptation (UDA) addresses this challenge by transferring knowledge from a labeled source domain to an unlabeled target domain to reduce the annotation cost. Significant research efforts have been devoted to developing effective techniques to minimize the domain gap between source and target domains. Recently, Vision Transformer (ViT) [5] has demonstrated great performance in domain adaptation tasks due to its superior generalization ability and has garnered significant research interest [23, 30, 31, 33]. State-of-the-art method PMTrans [33] proposes to construct an intermediate domain with mixup at patch level to bridge the domain gap and achieves superior performance. However, on the most challenging DomainNet [16] dataset, which exhibits a significant domain gap, PMTrans achieves an average accuracy of 52.4%, which indicates that there is still ample room for improvement.

How to address the adaptation problem when domain gap becomes large is an important research problem in domain adaptation. Existing methods [9, 17] leverage masked image modeling [8, 29], which is an effective self-supervised learning technique, with consistency regularization to improve target domain performance and demonstrate promising results, where they constrain the prediction of masked target data to be consistent with that of original target

data. However, existing methods simply treat masked image modeling as a form of data augmentation for regularization at the data instance level. They fail to consider the problem at a more grand distribution level, where all masked data instances form a "masked domain" that is intimately connected to the original domain. As shown in Fig. 1, the domain gap between masked domain and original domain is overlooked and not minimized by existing methods, which undermines domain adaptation performance when domain gap becomes large.

In this paper, we propose Dual Masked Generative Adversarial Transformer (**Dual-MGAT**), which leverages masked image modeling to construct masked domains with dual masked generative adversarial adaptation to fully unleash the potential of both transformer and masked image modeling to address the large domain gap problem in domain adaptation. Specifically, we formulate a general framework for masked domain learning. We assign pseudo-labels to original target data, and select high quality pseudo-labels as real labels for corresponding masked target data for learning. To reduce the domain gap between masked domain (e.g. masked target) and original domain (e.g. source), we realize that there exists two types of discrepancies: (1) information gap, where masked data contains only partial information due to masked image modeling; (2) domain distribution discrepancy, where masked data are constructed from a different domain compared to the original domain. To this end, we propose the masked generative adversarial adaptation technique, which introduces a transformer decoder to bridge the information gap between the masked feature space and the original feature space via [CLS] token feature generation and utilizes a class conditional domain discriminator for adversarial distribution alignment to reduce the domain distribution discrepancy. Our technique leads to much better performance than existing adversarial adaptation techniques that naively align the masked domain towards original domain without considering the two types of discrepancies. We further investigate the dual branch of masked source to target domain adaptation, which boosts model performance. Finally, we provide a theoretical analysis of our framework.

In summary, we have made following contributions in this paper:

- We propose dual masked generative adversarial transformer, which is an integrated framework to address the large domain gap problem with masked image modeling.
- We study masked domain adaptation and propose masked generative adversarial adaptation technique, which simultaneously resolve both information gap and domain distribution discrepancy at [CLS] token feature space and significantly outperforms existing adversarial adaptation techniques that naively align masked and original domain.
- We investigate the dual branch of masked source to target domain adaptation, which boosts performance.
- We provide a theoretical analysis of our framework from a masked domain adaptation perspective, which can help as guidance for future studies in leveraging masked image modeling for domain adaptation.
- We empirically validate the superiority of our framework with extensive experimental studies on three public domain adaptation benchmark datasets

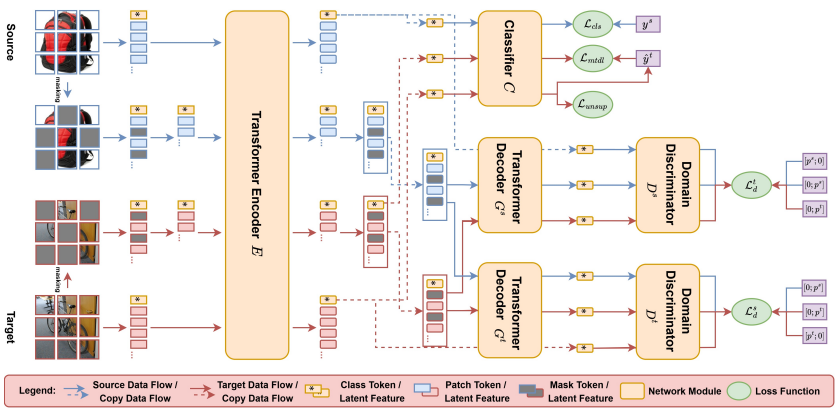

Fig. 2: Overview of our proposed dual masked generative adversarial transformer (**Dual-MGAT**) framework. Our framework consists of a transformer encoder, a classifier, two transformer decoders, and two class-conditional domain discriminators. Our framework constructs both masked source and masked target domains with masked image modeling to assist domain adaptation. Pseudo-labels on original target data $\hat{y}^t$ are leveraged for masked domain learning. Two pairs of transformer decoder and class-conditional domain discriminator, namely (G^s, D^s) and (G^t, D^t) perform masked generative adversarial adaptation from masked target to source and masked source to target domain respectively with the transformer encoder. Predictions on original source and target data p^s and p^t are leveraged to train the class-conditional domain discriminators for adversarial adaptation.

and one medical domain adaptation task. Our method achieves the best performance across all the evaluation datasets.

2 Related Works

Unsupervised Domain Adaptation aims to transfer knowledge from a labeled source domain to an unlabeled target domain to reduce the annotation cost. Existing domain adaptation methods can be generally categorized into feature distribution alignment methods, adversarial domain adaptation methods, self-training based methods, and generative adversarial adaptation methods. Feature distribution alignment methods reduce the distribution discrepancy across domains via minimizing certain discrepancy measures, such as Maximum Mean Discrepancy (MMD) [12,24], correlation statistics [22], or prototype-based distances [28,34]. Adversarial domain adaptation methods [6] follow the Generative Adversarial Nets (GAN) [7] framework, where a domain discriminator is introduced to distinguish source and target data and the feature encoder is adversarially trained to confuse the domain discriminator. Fine grained adversarial adaptation introduces a class conditional domain discriminator to both

distinguish the domains and learn to model the class structure [26]. Self-training based methods [37] assign pseudo-labels to unlabeled target data with the source trained network for self-training. Generative adversarial adaptation methods build the whole generation pipeline for domain adaptation. CoGAN [10] learns to generate the joint distribution of multi-domain data and attaches a domain invariant classifier at the last layer of the discriminator. PixelDA [2] generates target-alike images from labeled source data for supervised learning a classification model for target domain. GTA [21] learns a shared joint embedding space for both classification and image generation while enforcing feature domain invariance. Existing generative adversarial adaptation methods have mainly focused on image generation at the pixel space with original images, which by itself is a challenging task. Thus, their methods have been limited to some simple tasks, like digits classification, and simple object classification. To the best of our knowledge, we are the first to investigate masked image modeling for [CLS] token feature generation at the feature space to bridge the domain gap.

Masked Image Modeling is an effective representation learning technique in self-supervised learning [8, 29, 36]. Inspired from the success of masked image modeling, PACMAC [17] proposes to adapt self-supervised transformer from a source domain to a target domain via probing attention-conditioned masking consistency to identify reliable candidates for self-training. MIC [9] proposes to enforce consistency between predictions of masked target images and pseudo-labels generated based on complete images for robust visual recognition to improve domain adaptation performance. Existing masked image modeling based methods have shown promising domain adaptation results. However, they simply treat masked image modeling as a form of data augmentation and fail to view the problem at a more grand distribution level, thus they fail to fully unleash the potential of masked image modeling for unsupervised domain adaptation.

Vision Transformer. Since its introduction, vision transformer (ViT) [5] has garnered significant interest in solving various computer vision tasks. Recently, several studies have explored transformer for unsupervised domain adaptation and have demonstrated great performance. CDTrans [30] proposes a cross-attention module with self-training for domain adaptation. TVT [31] leverages transferable attention mechanism for adversarial adaptation to reduce the domain gap. SSRT [23] adopts transformer as backbone and proposes safe refinement learning with perturbed target data to refine the model. PMTrans [33] proposes mixup at patch-level to learn with an intermediate domain for domain adaptation. Our previous work, Ad^{2}mix [35], studies adversarial and adaptive mixup for UDA, where source and target samples are interpolated to construct mixup-based intermediate domains and adaptive mixup coefficients are used to create a natural curriculum for gradual adaptation. Different from these mixup-based intermediate domain learning frameworks, this work investigates masked domain adaptation with transformer by constructing masked source and target domains through masked image modeling and performing dual masked generative adversarial adaptation between masked and original domains.

3 Method

In unsupervised domain adaptation, we are given N^s labeled data $\mathbb{D}^s = \{(\boldsymbol{x}_i^s, y_i^s)\}_{i=1}^{N^s}$ in source domain and N^t unlabeled data $\mathbb{D}^t = \{\boldsymbol{x}_i^t\}_{i=1}^{N^t}$ in target domain. The source and target data share the same set of M labels and are sampled from probability distributions P^s and P^t respectively with $P^s \neq P^t$. The goal is to learn a model with labeled source data and unlabeled target data that can perform well in target domain. In Fig. 2, we present an overview of our proposed dual masked generative adversarial transformer (**Dual-MGAT**) framework.

3.1 Masked Domain Learning

We formulate a general framework for masked domain learning with masked image modeling for domain adaptation. Denote $F : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^M$ as a deep neural network, which is composed of a transformer encoder $E : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{(1+L)\times Z}$, where L is the number of visible image patches and Z is the hidden feature dimension and a classifier $C : \mathbb{R}^Z \rightarrow \mathbb{R}^M$. The classifier C takes the [CLS] token feature as input for prediction, where $F = E[0] \circ C$ and $E[0]$ outputs the [CLS] token feature. We utilize the labeled source data to train the network to learn discriminative features for prediction. Inspired from [9,17], we propose to learn a masked target domain that is intimately connected to original target domain to assist domain adaptation. Specifically, we apply masked image modeling [8] on target images and utilize the pseudo-labels assigned to original target images as supervision to train with the corresponding masked target images. We utilize a threshold value η to select pseudo-labels whose maximum predicted probability is larger than η for learning. Learning with the labeled source data and pseudo-labeled masked target data, the network features will tend to be biased towards them, which is however not desirable as the learned features may not be discriminative for target data. To this end, we further introduce an unsupervised learning term to minimize the entropy on target data to enforce the learned features are discriminative for target data. The objective of our masked target domain learning framework is defined as follows:

$$\mathcal{L}_{cls}(C, E) = \mathbb{E}[\mathcal{L}_{ce}(F(\boldsymbol{x}_i^s), y_i^s)], \tag{1}$$

$$\mathcal{L}_{mdl}(C, E) = \mathbb{E}[\mathbb{1}[\max(F(\boldsymbol{x}_i^t)) > \eta]\mathcal{L}_{ce}(F(\mathcal{M}[\boldsymbol{x}_i^t, \rho]), \hat{y}_i^t)], \tag{2}$$

$$\mathcal{L}_{unsup}(C, E) = H(\mathbb{E}[F(\boldsymbol{x}_i^t)]) - \mathbb{E}[H(F(\boldsymbol{x}_i^t))], \tag{3}$$

$$\mathcal{L}_{mtdl} = \mathcal{L}_{cls} + \mathcal{L}_{mdl} + \lambda_{unsup}\mathcal{L}_{unsup}. \tag{4}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss, $\mathbb{1}[\cdot]$ is the indicator function, $Max(\cdot)$ outputs the maximum value of a vector, $\mathcal{M}[\cdot, \rho]$ performs the masking operation with mask ratio ρ on an image, $\hat{y}_i^t = argmax F(\boldsymbol{x}_i^t)$ is the pseudo-label assigned to a target image, H calculates the entropy for each prediction probability, and λ_{unsup} is the balancing weight which is empirically set to be 0.01. Note, in the unsupervised learning loss, we also maximize global entropy to ensure all target data is not assigned to the same class.

3.2 Masked Generative Adversarial Adaptation

Masked target domain learning learns a masked target domain that is intimately connected to original target domain to facilitate knowledge transfer across source and target domain. However, as illustrated in Fig. 1, the domain gap between the masked target domain and source domain is non-negligible and will likely hinder the domain adaptation performance, especially when the domain gap becomes large. To this end, we propose to study the task of **masked domain adaptation** to reduce the domain gap between the masked target domain and source domain, which is largely overlooked by existing methods [9,17]. One naive solution to the task is to treat masked target domain as an independent domain and forcefully align the masked target domain towards source domain using existing domain adaptation techniques. However, such strategy ignores the information gap between masked domain and the original domain, which fails to effectively address the intricate discrepancies between masked target and source domain. Specifically, there exists two types of discrepancies: (1) information gap due to masked image modeling; (2) domain distribution discrepancy due to disparate data distribution between target and source domain. To this end, we propose to handle each type of discrepancy separately in order to fully resolve the domain gap between the masked target and source domain.

Specifically, we propose to learn the transformer encoder and a transformer decoder $G^s : \mathbb{R}^{(1+O)\times Z} \rightarrow \mathbb{R}^{(1+O)\times Z}$, where O is the number of patches for a complete image, to reconstruct the original source data from masked source data to bridge the information gap. Note, different from masked autoencoder [8], which reconstructs the complete images from masked images at the pixel space, we propose to reconstruct the [CLS] token features of complete images from that of masked images at the feature space. We argue that reconstructing at pixel space requires the network to learn features concerning low level image details, which may not be useful to our task. On the contrary, [CLS] token feature is the global representation of the whole image and contains rich semantic information.

Following [8], we input the encoded visible latent features together with the mask latent features to the transformer decoder. Each mask latent feature is a shared, learnable vector that indicates the presence of a missing patch. We add positional embedding to all latent features to distinguish mask latent features at different positions. To reconstruct [CLS] token feature, we propose to utilize the less restrictive adversarial training loss [7] instead of the commonly adopted element-wise l^2 loss for the task [8, 29], where we think too restrictive reconstruction loss may adversely affect feature representation learning. We provide empirical comparison results in Sec. 4.3. We adopt a class-conditional domain discriminator $D^s : \mathbb{R}^Z \rightarrow \mathbb{R}^{2\times M}$ to both encode the domain and class information to distinguish the reconstructed masked source [CLS] token feature and the original source [CLS] token feature and adversarially train the transformer encoder and decoder to ensure the two types of [CLS] token features are indistinguishable.

To reduce the domain distribution discrepancy, we note that the transformer decoder is trained to reconstruct the masked source [CLS] token feature to-

wards the original source [CLS] token feature. If there is no domain distribution discrepancy between target and source domain, then the reconstructed masked target [CLS] token feature with the same transformer decoder will be alike reconstructed masked source [CLS] token feature and be aligned towards the original source [CLS] token feature and vice versa. Therefore, we propose to align the reconstructed masked target [CLS] token feature towards the original source [CLS] token feature to reduce the domain distribution discrepancy. We utilize the same domain discriminator D^s in [CLS] token feature reconstruction to distinguish the reconstructed masked target [CLS] token feature from original source [CLS] token feature and adversarially train the transformer encoder to align the masked target feature distribution towards masked source feature distribution so that the transformer decoder can reconstruct the masked target [CLS] token feature similarly as that of masked source data. For both [CLS] token feature reconstruction and adversarial distribution alignment, we utilize network predictions on original source and target data to encode semantic information and concatenate them with $\mathbf{0}$ vectors to encode domain information as the targets for training. The transformer encoder, transformer decoder, and domain discriminator play a three-player min-max game following the GAN framework [7]. At the optimal, the transformer encoder and transformer decoder learn to bridge both the information gap and domain distribution discrepancy so that the domain discriminator cannot distinguish the reconstructed masked source and masked target [CLS] token feature from original source [CLS] token feature anymore. We term our technique in reducing the domain gap between masked target and source domain as masked generative adversarial adaptation. The training loss function is defined as follows:

$$\begin{aligned}\mathcal{L}_d^s(D^s) = &\mathbb{E}[\mathcal{L}_{ce}(D^s(E(\boldsymbol{x}_i^s)[0]), [\boldsymbol{p}_i^s; \mathbf{0}])] \\ &+ \mathbb{E}[\mathcal{L}_{ce}(D^s(G^s(\boldsymbol{f}_i^s)[0]), [\mathbf{0}; \boldsymbol{p}_i^s])] \\ &+ \mathbb{E}\big[\mathcal{L}_{ce}\big(D^s(G^s(\boldsymbol{f}_i^t)[0]), [\mathbf{0}; \boldsymbol{p}_i^t]\big)\big]\,, \end{aligned} \tag{5}$$

$$\mathcal{L}_{adv,i}^s(G^s, E) = \mathbb{E}[\mathcal{L}_{ce}(D^s(G^s(\boldsymbol{f}_i^s)[0]), [\boldsymbol{p}_i^s; \mathbf{0}])]\,, \tag{6}$$

$$\mathcal{L}_{adv,d}^s(E) = \mathbb{E}\big[\mathcal{L}_{ce}\big(D^s(G^s(\boldsymbol{f}_i^t)[0]), [\boldsymbol{p}_i^t; \mathbf{0}]\big)\big]\,, \tag{7}$$

$$\mathcal{L}_{mgaa}^s = \mathcal{L}_d^s + \lambda_{adv}\big(\mathcal{L}_{adv,i}^s + \mathcal{L}_{adv,d}^s\big). \tag{8}$$

where $\boldsymbol{f}_i^s = [E(\mathcal{M}[\boldsymbol{x}_i^s, \rho]); \boldsymbol{T}]$, $\boldsymbol{T}$ denotes the set of mask latent features with positional embedding for missing patches, $[\cdot;\cdot]$ concatenates two input vectors into a single vector, $\boldsymbol{f}_i^t = [E(\mathcal{M}[\boldsymbol{x}_i^t, \rho]); \boldsymbol{T}]$, $\boldsymbol{p}_i^s = F(\boldsymbol{x}_i^s)$, $\boldsymbol{p}_i^t = F(\boldsymbol{x}_i^t)$, $\mathbf{0}$ is a M-dimensional zero vector, and λ_{adv} is the balancing weights which is empirically set to be 0.1.

3.3 Extension to the Dual Branch

Masked target to source domain adaptation reduces the domain gap between masked target and source domain, which helps to align target domain towards source domain. However, such alignment is one-directional and as shown in

Fig. 1, the information gap between masked target and original target domain is not minimized. To this end, we propose to further impose the dual branch of masked source to target domain adaptation into our framework to (1) bridge information gap between masked target and original target domain; (2) form a comprehensive bidirectional alignment path to more effectively reduce domain gap between source and target domain. Thus, we introduce another pair of transformer decoder G^t and domain discriminator D^t and define the training loss $\mathcal{L}^t_{mgaa}$ accordingly based on the proposed masked generative adversarial adaptation technique.

3.4 Overall Objective

The overall objective for our dual masked generative adversarial transformer (**Dual-MGAT**) framework is defined as follows:

$$\min_{E,C,G^s,D^s,G^t,D^t} \mathcal{L}_{mtdl} + \mathcal{L}^s_{mgaa} + \mathcal{L}^t_{mgaa}. \tag{9}$$

3.5 Theoretical Analysis

Inspired from [1], we treat masked target domain as a noisy labeled domain and show that our framework is minimizing an upper bound on the target expectation error as below:

Theorem 1 (Masked Domain Adaptation Theorem). *Following our problem setting, let h be a hypothesis in class $\mathcal{H}$, let γ be the noisy ratio of pseudo-labeled masked target domain, denote the probability distribution of masked target domain as $P^{t'}$, then for any $\delta \in (0,1)$, with probability at least $1-\delta$, for every $h \in \mathcal{H}$, we have:*

$$\begin{aligned} \epsilon_t(h) \leq & \frac{1}{2}\epsilon_s(h) + \frac{1}{2}\epsilon_{t'}(h) + \frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(P^{t'}, P^s) \\ & + \frac{1}{4}d_{\mathcal{H}\Delta\mathcal{H}}(P^{t'}, P^t) + 2\gamma + C^*_1 + C^*_2, \end{aligned} \tag{10}$$

where $\epsilon_t(\cdot)$ (resp. $\epsilon_{t'}(\cdot)$, $\epsilon_s(\cdot)$) measures the expectation error of a hypothesis on target (resp. masked target, source) data distribution, $d_{\mathcal{H}\Delta\mathcal{H}}(\cdot,\cdot)$ measures the distribution discrepancy between two data distributions, $C^*_1 = \min_{h\in\mathcal{H}} \epsilon_{t'}(h) + \epsilon_t(h)$, and $C^*_2 = \min_{h\in\mathcal{H}} \epsilon_{t'}(h) + \epsilon_s(h)$.

The theorem upper bounds the target expectation error with (1) source expectation error; (2) pseudo-labeled masked target expectation error; (3) the domain gap between masked target and source domain; (4) the domain gap between masked target and target domain; (5) pseudo-label noisy ratio; and (6) the non-optimizable optimal error between masked target and target domain and masked target and source domain that are assumed to be small [1]. Our masked target domain learning loss minimizes terms (1), (2) and we apply confidence threshold to reduce term (5). The masked generative adversarial adaptation from masked

Table 1: Effectiveness of masked generative adversarial adaptation (MGAA) on DomainNet. The best performance is marked as **bold**.

Method	clp→inf	inf→clp	Avg
Source Only	29.5	61.3	45.4
+FADA [26]	33.1	67.9	50.5
+**MGAA**	**33.7**	**69.6**	**51.7**
MTDL	34.8	73.6	54.2
+FADA [26]	35.0	73.9	54.5
+**MGAA** (reduce info. gap only)	35.1	74.2	54.7
+**MGAA**	**35.8**	**74.8**	**55.3**

Table 2: Effectiveness of [CLS] token feature reconstruction and adversarial training loss on DomainNet under source supervised learning. The best performance is marked as **bold**.

Pixel	[CLS]	l^2 Loss	Adv. Loss	clp→inf	inf→clp	Avg
✓		✓		27.2	61.0	44.1
✓			✓	32.4	62.7	47.6
	✓	✓		32.7	66.7	49.7
	✓		✓	**33.7**	**69.6**	**51.7**

target to source domain and masked source to target domain minimize terms (3), (4) respectively. Existing masked image modeling based adaptation methods [9,17] can be regarded as approximately minimizing terms (1), (2), and (5) as masked target domain learning. Thus, they demonstrate effectiveness in domain adaptation, however, they overlook and fail to minimize terms (3) and (4).

4 Experiments

4.1 Datasets and experiment setting

Datasets. We evaluate our framework on three public domain adaptation benchmark datasets and one medical domain adaptation task. **Office-31** [19] consists of 4,652 images of 31 classes from three domains: Amazon (A), DSLR (D), and Webcam (W). **Office-Home** [25] contains 15,500 images of 65 classes from four domains: Artistic (A), Clipart (C), Product (P), and Real-world (R). **DomainNet** [16] is a challenging domain adaptation dataset with large domain gap, which contains about 0.6 million images of 345 classes in six domains: Clipart (clp), Infograph (inf), Painting (pnt), Quickdraw (qdr), Real (rel), and Sketch (skt). **Cataract** consists of 1,428 eye images of 2 classes collected with two different devices: Slit-Lamp (S) and Camera (C). The Slit-Lamp dataset is a private dataset of eye images captured with specialized slit-lamp device in a local hospital for cataract detection. The Camera dataset[4] is a public Kaggle competition dataset of eye images captured with camera device. Due to the nature of different devices and procedures for eye image capturing, the two domains have large domain gap. Due to space limit, we display images in each domain and present more details on the dataset in Appendix.

Implementation. In all experiments, we adopt the Vision Transformer (ViT) base model [5] pre-trained on ImageNet [4] as the feature encoder. We implement the classifier as a two-layer multi-layer perceptron (MLP) with GELU non-linearity. We implement the decoder as a transformer with four blocks. The domain discriminator is implemented as a MLP with the same architecture as the classifier except for the output. We train our framework with a base learning

[4] https://www.kaggle.com/datasets/nandanp6/cataract-image-dataset

Table 3: Ablation Study. The best performance is marked as **bold**.

$\mathcal{L}_{cls}$	$\mathcal{L}_{mtdl}$	$\mathcal{L}^s_{mgaa}$	$\mathcal{L}^t_{mgaa}$	clp→inf	inf→clp	Avg
✓				29.5	61.3	45.4
✓		✓		33.0	68.4	50.7
✓			✓	33.7	69.6	51.7
✓		✓	✓	34.5	71.2	52.9
	✓			34.8	73.6	54.2
	✓	✓		35.5	74.3	54.9
	✓		✓	35.8	74.8	55.3
	✓	✓	✓	**36.4**	**75.6**	**56.0**

Table 4: Comparison with state-of-the-art methods on Cataract. The best performance is marked as **bold**.

Method	Slit-Lamp→Camera		
	AUC (%)	Sensitivity (%)	Specificity (%)
Deit	64.1	57.6	74.8
CDTrans	73.7	72.5	76.5
ViT-B	71.6	58.3	76.7
SSRT	72.3	63.4	77.8
TVT	80.9	73.2	79.1
PMTrans	85.9	73.7	85.7
Dual-MGAT	**94.8**	**78.8**	**97.1**

rate of 1e-5 and a batch size of 32 for 50 epochs. We set the learning rate of the classifier to be 10 times higher as the base learning rate. We adopt AdamW [13] with a momentum of 0.9 and a weight decay of 0.05 as the optimizer. The mask ratio ρ and the threshold η are set to 0.7 and 0.7 respectively for all experiments.

Comparison Methods. We compare our method with state-of-the-art domain adaptation methods, including both ResNet-based and transformer-based methods. The ResNet-based methods include FixBi [15], MCD [20], SCDA [11], BNM [3], SDAT [18], MDD [32], and GSDE [27]. The transformer-based methods include CDTrans [30], TVT [31], SSRT [23], MIC [9], PMTrans [33], and DPP&BST [14].

4.2 Effectiveness of Masked Generative Adversarial Adaptation

We study the effectiveness of our proposed masked generative adversarial adaptation (MGAA) technique for masked target to source domain adaptation under both source supervised learning setting and masked target domain learning (MTDL) setting. We compare our method with existing fine-grained class conditional adversarial adaptation method FADA [26], which directly align the masked target domain towards source domain. We also experiment with MGAA (reduce info. gap only), which only reduce the information gap between masked source and original source domain. Table 1 shows the experiment results. We have the following observations: (1) both FADA and MGAA improve over the baseline methods in both settings. **The results reveal that: (a) reducing the domain gap between masked target and source domain can help reduce the domain gap between source and target domain; (b) the domain gap between masked target and source domain can negatively affect masked target domain learning performance.** The results empirically confirm the necessity of our study of masked domain adaptation, which is largely overlooked by existing methods; (2) MGAA (reduce info. gap only) improves MTDL but underperforms MGAA. This is as expected, as reducing the information gap helps reduce part of the domain gap between masked target and source domain and further reducing domain distribution discrepancy helps to reduce all; (3) MGAA significantly outperforms FADA in both settings, which empirically validates that our strategy of handling each type of discrepancy sepa-

Table 5: Comparison with SoTA methods on Office-Home. The best performance is marked as **bold**.

Method		A→C	A→P	A→R	C→A	C→P	C→R	P→A	P→C	P→R	R→A	R→C	R→P	Avg
ResNet-50	ResNet	44.9	66.3	74.3	51.8	61.9	63.6	52.4	39.1	71.2	63.8	45.9	77.2	59.4
MCD	ResNet	48.9	68.3	74.6	61.3	67.6	68.8	57.0	47.1	75.1	69.1	52.2	79.6	64.1
MDD	ResNet	54.9	73.7	77.8	60.0	71.4	71.8	61.2	53.6	78.1	72.5	60.2	82.3	68.1
BNM	ResNet	56.7	77.5	81.0	67.3	76.3	77.1	65.3	55.1	82.0	73.6	57.0	84.3	71.1
SDAT	ResNet	58.2	77.1	82.2	66.3	77.6	76.8	63.3	57.0	82.2	74.9	64.7	86.0	72.2
FixBi	ResNet	58.1	77.3	80.4	67.7	79.5	78.1	65.8	57.9	81.7	76.4	62.9	86.7	72.7
GSDE	ResNet	57.8	80.2	81.9	71.3	78.9	80.5	67.4	57.2	84.0	76.1	62.5	85.7	73.6
Deit	ViT	61.8	79.5	84.3	75.4	78.8	81.2	72.8	55.7	84.4	78.3	59.3	86.0	74.8
CDTrans	ViT	68.8	85.0	86.9	81.5	87.1	87.3	79.6	63.3	88.2	82.0	66.0	90.6	80.5
DPP&BST	ViT	69.8	84.7	87.0	80.5	86.8	86.5	78.3	65.1	88.1	81.8	66.9	90.1	80.6
ViT-B	ViT	67.0	85.7	88.1	80.1	84.1	86.7	79.5	67.0	89.4	83.6	70.2	91.2	81.1
TVT	ViT	74.9	86.8	89.5	82.8	88.0	88.3	79.8	71.9	90.1	85.5	74.6	90.6	83.6
SSRT	ViT	75.2	89.0	91.1	85.1	88.3	89.9	85.0	74.2	91.2	85.7	78.6	91.8	85.4
MIC(SDAT)	ViT	80.2	87.3	91.1	87.2	90.0	90.1	83.4	75.6	91.2	88.6	78.7	91.4	86.2
Dual-MGAT	ViT	**82.8**	**90.7**	**92.4**	**89.0**	**92.5**	**92.7**	**87.7**	**81.8**	**92.7**	**90.0**	**82.9**	**94.2**	**89.1**

Table 6: Comparison with state-of-the-art methods on Office-31.

Method		A→W	D→W	W→D	A→D	D→A	W→A	Avg
ResNet-50	ResNet	68.9	68.4	62.5	96.7	60.7	99.3	76.1
BNM	ResNet	91.5	98.5	**100.0**	90.3	70.9	71.6	87.1
MDD	ResNet	94.5	98.4	**100.0**	93.5	74.6	72.2	88.9
SCDA	ResNet	94.2	98.7	99.8	95.2	75.7	76.2	90.0
FixBi	ResNet	96.1	99.3	**100.0**	95.0	78.7	79.4	91.4
GSDE	ResNet	96.9	98.8	**100.0**	96.7	78.3	79.2	91.7
Deit	ViT	89.2	98.9	**100.0**	88.7	80.1	79.8	89.5
CDTrans	ViT	96.7	99.0	**100.0**	97.0	81.1	81.9	92.6
ViT-B	ViT	91.2	99.2	**100.0**	90.4	81.1	80.6	91.1
TVT	ViT	96.4	99.4	**100.0**	96.4	84.9	86.0	93.9
SSRT	ViT	97.7	99.2	**100.0**	**98.6**	83.5	82.2	93.5
Dual-MGAT	ViT	**98.0**	**99.9**	100.0	**98.6**	**87.0**	**87.1**	**95.1**

rately to fully resolve the domain gap between masked target and source domain is both viable and more effective.

4.3 Effectiveness of [CLS] Token Feature Reconstruction and Adversarial Training Loss

We study the effectiveness of our proposed [CLS] token feature reconstruction and adversarial training loss to bridge the information gap for masked target to source domain adaptation under source supervised learning setting. Similar observations are made under MTDL setting (Please view Appendix). We compare our strategy with pixel space reconstruction and element-wise l^2 loss. Due to space limit, we leave the implementation details for different comparison methods in Appendix. Table 2 shows the experiment results. We have the following observations: (1) [CLS] token feature reconstruction always outperforms pixel space reconstruction in bridging the information gap regardless of the reconstruction loss function used. The results empirically conform our analysis that reconstruction at pixel space will enforce the network to learn features concerning low level image details, which may not be useful to our task; (2) the adversarial training loss outperforms the l^2 loss for both pixel space or [CLS] token feature reconstruction, which indicates that a less restrictive reconstruction loss function is more beneficial to bridge the information gap; (3) the combination of pixel space reconstruction with l^2 loss gives even worse results than source supervised learning baseline with an average accuracy of 45.4%. We attribute the worse results to the restrictive pixel space reconstruction task, which adversely affects source supervised learning task for transferable feature representations. **To the best of our knowledge, we are the first to investigate [CLS] token feature reconstruction with masked image modeling to bridge the information gap for domain adaptation and our findings are important for future studies which may similarly want to leverage masked image modeling for domain adaptation.**

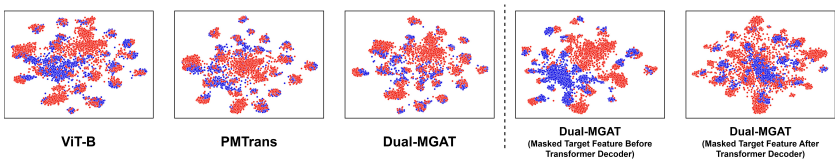

Fig. 3: T-SNE visualization of inf→clp transfer task with twenty-five classes on DomainNet. Source and target instances are shown in red and blue colors respectively.

Table 7: Comparison with SoTA methods on DomainNet. The best performance is marked as **bold**.

MDD	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	20.5	40.7	6.2	52.5	42.1	32.4
inf	33.0	–	33.8	2.6	46.2	24.5	28.0
pnt	43.7	20.4	–	2.8	51.2	41.7	32.0
qdr	18.4	3.0	8.1	–	12.9	11.8	10.8
rel	52.8	21.6	47.8	4.2	–	41.2	33.5
skt	54.3	17.5	43.1	5.7	54.2	–	35.0
Avg	40.4	16.6	34.7	4.3	43.4	32.3	28.6

SCDA	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	18.6	39.3	5.1	55.0	44.1	32.4
inf	29.6	–	34.0	1.4	46.3	25.4	27.3
pnt	44.1	19.0	–	2.6	56.2	42.0	32.8
qdr	30.0	4.9	15.0	–	25.4	19.8	19.0
rel	54.0	22.5	51.9	2.3	–	42.5	34.6
skt	55.6	18.5	44.7	6.4	53.2	–	35.7
Avg	42.6	16.7	37.0	3.6	47.2	34.8	30.3

CDTrans	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	29.4	57.2	26.0	72.6	58.1	48.7
inf	57.0	–	54.4	12.8	69.5	48.4	48.4
pnt	62.9	27.4	–	15.8	72.1	53.9	46.4
qdr	44.6	8.9	29.0	–	42.6	28.5	30.7
rel	66.2	31.0	61.5	16.2	–	52.9	45.6
skt	69.0	29.6	59.0	27.2	72.5	–	51.5
Avg	59.9	25.3	52.2	19.6	65.9	48.4	45.2

SSRT	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	33.8	60.2	19.4	75.8	59.8	49.8
inf	55.5	–	54.0	9.0	68.2	44.7	46.3
pnt	61.7	28.5	–	8.4	71.4	55.2	45.0
qdr	42.5	8.8	24.2	–	37.6	33.6	29.3
rel	69.9	37.1	66.0	10.1	–	58.9	48.4
skt	70.6	32.8	62.2	21.7	73.2	–	52.1
Avg	60.0	28.2	53.3	13.7	65.3	50.4	45.2

PMTrans	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	34.2	62.7	32.5	79.3	63.7	54.5
inf	67.4	–	61.1	22.2	78.0	57.6	57.3
pnt	69.7	33.5	–	23.9	79.8	61.2	53.6
qdr	54.6	17.4	38.9	–	49.5	41.0	40.3
rel	74.1	35.3	70.0	25.4	–	61.1	53.2
skt	73.8	33.0	62.6	30.9	77.5	–	55.6
Avg	67.9	30.7	59.1	27.0	72.8	56.9	52.4

Dual-MGAT	clp	inf	pnt	qdr	rel	skt	Avg
clp	–	36.4	68.3	37.6	80.3	68.1	58.1
inf	75.6	–	67.7	24.3	81.3	65.5	62.9
pnt	75.7	37.3	–	27.2	80.8	66.3	57.4
qdr	69.6	27.7	50.9	–	61.4	57.0	53.3
rel	78.4	37.7	71.1	29.9	–	67.1	56.8
skt	79.1	38.2	69.2	35.8	81.2	–	60.7
Avg	75.7	35.5	65.4	31.0	77.0	64.8	**58.2**

4.4 Ablation Study

We perform ablation study on our framework to investigate the joint effects of different components. Table 3 shows the experiment results. As can be observed, source only method performs poorly. Masked target domain learning, masked generative target to source adversarial adaptation, and masked generative source to target adversarial adaptation all can improve the domain adaptation performance, which indicates the effectiveness of each component. We further observe that the combination of any two of these three components improves the performance over each individual and the combination of all the three components achieves the best performance, which indicates the complementary roles of them played in boosting domain adaptation performance.

4.5 Comparison with State-of-the-Art Methods

In Tabs. 4 to 7, we evaluate the performance of our framework on three public domain adaptation benchmark datasets and one medical task on domain adaptive cataract detection. We observe that transformer-based methods generally outperform ResNet-based methods due to the superior generalizability of vision transformer. When compared to the baseline ViT-B method, our framework

improves over the baseline by **8.0%** and **4.0%** for Office-Home and Office-31 respectively. In Office-Home, our method outperforms state-of-the-art MIC(SDAT) method on all twelve transfer tasks among four domains and improves over it by **2.9%** on average accuracy. In Office-31, our method achieves an average accuracy result of **95.1%** over six transfer tasks among three domains, and outperforms other state-of-the-art UDA methods significantly. To the best of our knowledge, this is currently the best result on Office-31 with ViT-B. In the most challenging DomainNet dataset, our method outperforms existing methods significantly. Specifically, our method outperforms the previous best method PMTrans by **5.8%** on average accuracy result over 30 transfer tasks among six domains. Incredibly, our method outperforms PMTrans in all the 30 transfer tasks. On multiple challenging transfer tasks, for example, qdr→inf and qdr→pnt, our method has improved over PMTrans by **10.3%** and **12.0%** respectively, which are tremendous improvements and highlight the strong ability of our method in handling datasets with large domain gap. In Cataract, our method achieves significantly better AUC, sensitivity, and specificity results than state-of-the-art domain adaptation methods, which demonstrates the superiority of our method in handling domain shift in a real-world application scenario.

4.6 Feature Visualization

Fig. 3 presents the T-SNE visualization results of the ViT-B baseline method, the PMTrans method, and our proposed Dual-MGAT method on inf→clp transfer task on DomainNet. As can be observed, both our method and PMTrans better align target features towards source domain when compared to ViT-B baseline and our method demonstrates slightly better qualitative alignment result than PMTrans. In column 4 and 5, we display the masked target feature before and after the transformer decoder. We observe that there exists apparent domain gap between masked target feature and source feature before the transformer decoder, which is largely due to information gap. With the transformer decoder, the masked target feature shows qualitatively better alignment results towards the source domain, which validates the effectiveness of our design of masked generative adversarial adaptation at the [CLS] token feature space.

4.7 More Experimental Analysis

Due to space limit, we put more experiment results on the effect of **decoder depth**, **mask ratio**, and **confidence threshold** in Appendix. Experiment results show that more depth gives slightly better performance but plateau at 4. Mask ratio affects information gap between masked and original domain. Too small or too large value gives slightly poorer performance and the default value of 0.7 gives the best performance. Masked image modeling mitigates the learning of false correlations, enhancing our framework's robustness to confidence threshold. We also conduct sensitivity analysis on hyper-parameters λ_{adv} and λ_{unsup} in Appendix, showing that our method is robust to hyper-parameter change in a wide

range. We further report computational cost in Appendix, where Dual-MGAT incurs only modest training overhead and no additional inference cost.

5 Conclusion

In this paper, we present a dual masked generative adversarial transformer framework for UDA. We study masked domain adaptation and propose masked generative adversarial adaptation technique at [CLS] token feature space, which simultaneously reduces both the information gap and domain distribution discrepancy across masked target and source domain and shows better adaptation performance than existing adversarial adaptation technique. We further investigate the dual branch of masked source to target domain adaptation, which boosts model performance. We provide a theoretical analysis of our framework from a masked domain adaptation perspective, which highlights the essentials in leveraging masked image modeling for domain adaptation. Extensive experimental studies on three public datasets and one medical adaptation dataset demonstrate the superiority of our framework. We hope our investigation towards masked image modeling for domain adaptation can inspire future studies to not merely treat masked image modeling as a form of data augmentation for consistency regularization at data instance level and can pay more attention to the often ignored domain gap between masked and original domain.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., Vaughan, J.W.: A theory of learning from different domains. Machine learning **79**(1-2), 151–175 (2010) 9
2. Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., Krishnan, D.: Unsupervised pixel-level domain adaptation with generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3722–3731 (2017) 5
3. Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q., Tian, Q.: Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3941–3950 (2020) 11
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 10
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy 2, 5, 10
6. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. p. 1180–1189. ICML'15, JMLR.org (2015) 4

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014) 4, 7, 8
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022) 2, 5, 6, 7
9. Hoyer, L., Dai, D., Wang, H., Van Gool, L.: Mic: Masked image consistency for context-enhanced domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11721–11732 (2023) 2, 5, 6, 7, 10, 11
10. Li, Q., Lu, L., Li, Z., Wu, W., Liu, Z., Jeon, G., Yang, X.: Coupled gan with relativistic discriminators for infrared and visible images fusion. IEEE Sensors Journal **21**(6), 7458–7467 (2019) 5
11. Li, S., Xie, M., Lv, F., Liu, C.H., Liang, J., Qin, C., Li, W.: Semantic concentration for domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9102–9111 (2021) 11
12. Long, M., Wang, J., Ding, G., Sun, J., Yu, P.S.: Transfer joint matching for unsupervised domain adaptation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1410–1417 (2014) 4
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2017), https://api.semanticscholar.org/CorpusID:53592270 11
14. Luo, H., Tian, Z., Jiao, P., Liu, M., Du, S., Nan, K.: Explicitly fusing plug-and-play guidance of source prototype into target subspace for domain adaptation. Information Fusion **123**, 103197 (2025) 11
15. Na, J., Jung, H., Chang, H.J., Hwang, W.: Fixbi: Bridging domain spaces for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1094–1103 (2021) 11
16. Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., Wang, B.: Moment matching for multi-source domain adaptation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1406–1415 (2019) 2, 10
17. Prabhu, V., Yenamandra, S., Singh, A., Hoffman, J.: Adapting self-supervised vision transformers by probing attention-conditioned masking consistency. Advances in Neural Information Processing Systems **35**, 23271–23283 (2022) 2, 5, 6, 7, 10
18. Rangwani, H., Aithal, S.K., Mishra, M., Jain, A., Radhakrishnan, V.B.: A closer look at smoothness in domain adversarial training. In: International conference on machine learning. pp. 18378–18399. PMLR (2022) 11
19. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting visual category models to new domains. In: European conference on computer vision. pp. 213–226. Springer (2010) 10
20. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3723–3732 (2018) 11
21. Sankaranarayanan, S., Balaji, Y., Castillo, C.D., Chellappa, R.: Generate to adapt: Aligning domains using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8503–8512 (2018) 5
22. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European Conference on Computer Vision. pp. 443–450. Springer (2016) 4

23. Sun, T., Lu, C., Zhang, T., Ling, H.: Safe self-refinement for transformer-based domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7191–7200 (2022) 2, 5, 11
24. Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., Darrell, T.: Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474 (2014) 4
25. Venkateswara, H., Eusebio, J., Chakraborty, S., Panchanathan, S.: Deep hashing network for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5018–5027 (2017) 10
26. Wang, H., Shen, T., Zhang, W., Duan, L.Y., Mei, T.: Classes matter: A fine-grained adversarial approach to cross-domain semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV. pp. 642–659. Springer (2020) 5, 10, 11
27. Westfechtel, T., Yeh, H.W., Zhang, D., Harada, T.: Gradual source domain expansion for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1946–1955 (2024) 11
28. Xie, S., Zheng, Z., Chen, L., Chen, C.: Learning semantic representations for unsupervised domain adaptation. In: International Conference on Machine Learning. pp. 5423–5432 (2018) 4
29. Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H.: Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663 (2022) 2, 5, 7
30. Xu, T., Chen, W., WANG, P., Wang, F., Li, H., Jin, R.: CDTrans: Cross-domain transformer for unsupervised domain adaptation. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=XGzk5OKWFFc 2, 5, 11
31. Yang, J., Liu, J., Xu, N., Huang, J.: Tvt: Transferable vision transformer for unsupervised domain adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 520–530 (2023) 2, 5, 11
32. Zhang, Y., Liu, T., Long, M., Jordan, M.: Bridging theory and algorithm for domain adaptation. In: International conference on machine learning. pp. 7404–7413. PMLR (2019) 11
33. Zhu, J., Bai, H., Wang, L.: Patch-mix transformer for unsupervised domain adaptation: A game perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3561–3571 (2023) 2, 5, 11
34. Zhu, L., Luo, Z., Wang, W., Zhang, M., Chen, G., Zheng, K.: Towards robust cross-domain image understanding with unsupervised noise removal. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 3024–3033 (2021) 4
35. Zhu, L., Xu, Y., Liu, Y., Goh, R.S.M., Xu, X.: Ad2mix: Adversarial and adaptive mixup for unsupervised domain adaptation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 6581–6590. IEEE (2025) 5
36. Zhu, L., Zhou, J., Goh, R.S.M., Liu, Y.: Advmim: Adversarial masked image modeling for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 56–66. Springer (2025) 5
37. Zou, Y., Yu, Z., Kumar, B.V., Wang, J.: Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 289–305 (2018) 5